Examining Testing Policy in the United States: A Comparative Historical Analysis of National Testing for Accountability Debates and Intelligence Testing Debates

By
Gordon Thomas Way
University of Kansas

Submitted to the graduate degree program in Education Leadership and Policy Studies and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Education.

_____
Chairperson Dr. John Rury, Ph.D.

_____
Dr. Rick Ginsberg, Ph.D.

_____
Dr. Mickey Imber, Ph.D.

_____
Dr. Argun Saatcioglu, Ph. D.

_____
Dr. Marc Mahlios, Ph.D.

Date Defended: August 28, 2014

The Dissertation Committee for Gordon Thomas Way
certifies that this is the approved version of the following dissertation:

Examining Testing Policy in the United States: A Comparative Historical Analysis of National Testing for Accountability Debates and Intelligence Testing Debates

_____
Chairperson Dr. John Rury, Ph.D.

Date approved:

Abstract

Educational testing policies influence the type of education provided to school children, and communicate what society values. These policies originate from and promote a set of assumptions, beliefs, and values orientations. They are also designed to advance certain social and educational outcomes, and preferred types of educational experiences for school children. As lawmakers deliberate future testing policies in the United States, it is important that they understand the motivations and values behind testing policy proposals.

This study explores the beliefs that drove accountability testing policy proposals. It employed a historical/ comparative analysis to compare the arguments from the accountability testing debates of the 1980s, 1990s, and early 2000s with the arguments from the intelligence testing debates of the early 1920s, the 1970s, and mid-1990s in order to tease out the beliefs that drove these arguments. Issues around which the arguments revolved were examined to identify enduring themes that can inform future educational testing policy proposals.

Six such themes emerged from this analysis. These included the following: 1) merit; 2) "race," class and educational equity; 3) the meaning of democracy; 4) the fundamental purpose of public education and desired educational experiences in the United States; 5) the role of science and ideology in policy making; and 6) the tendency to oversimplify. These themes and their implications for policy were discussed.

Acknowledgements

I would like to thank Dr. John Rury, Ph.D., for consistently guiding me to write the kind of dissertation I wanted to write. In addition, I would like to thank my committee members Dr. Rick Ginsberg, Ph.D., Dr. Mickey Imber, Ph.D., Dr. Argun Saatcioglu, Ph.D., and Dr. Marc Mahlios, Ph.D., for holding me to a rigorous standard. Their feedback has made this a better dissertation.

Finally, I must acknowledge my wife Beatriz, and my children Isabella and Thomas for their patience and support in this process.

Table of Contents

Chapter 1: Introduction and research design

Introduction

With its warning 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," *A Nation at Risk: The Imperative for Educational Reform* (Commission for Excellence in Education, 1983, p. 7) launched a wave of public school criticism that included increased calls for holding educators and educational systems accountable for student test scores (Baker & Stites, 1991; Darling-Hammond & Berry 1988). This use of testing for accountability as a means to reform public education was codified into law with the 2002 reauthorization of the Elementary and Secondary Schools Act, commonly known as *The No Child Left Behind Act* (No Child Left Behind [NCLB], 2002).

NCLB (2002) mandated that students be tested annually, and that educators, schools, districts, and states be held accountable for the aggregate and disaggregated performance of students on these tests. The purpose was to use a system of support, rewards and sanctions to improve education for all students, including historically underserved poor, minority, disabled students, and limited English proficient students.

Although the stated purpose of NCLB (2002) was to raise achievement for all American students, the law was controversial. Critics claimed that the combination of standardized tests and high stakes for educators, and sometimes students, had serious unintended consequences. Among others, these consequences included a narrowing of the curriculum and educational experiences for students (Madaus, Russell & Higgins, 2009; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004), inflated test scores (Koretz, 2008; Madaus, Russell &

Higgins, 2009), the undermining of schools serving the neediest students (Madaus, Russell, & Higgins, 2009; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004), motivating states to lower proficiency standards (Cawelti, 2006; Fusarelli, 2004; Lewis, 2002; Peterson & Hess, 2007), and placing the impossibly high objective of reaching 100 percent student proficiency in reading and math by 2014 (Darling-Hammond, 2006; Haas, Wilson, Cobb, Rallis, 2005; Kohn, 2004a, 2004b; Ravitch, 2011).

Although it was originally due for reauthorization in 2007, congress has yet to pass an NCLB (2002) reauthorization bill (Duncan, July 19, 2013). As lawmakers deliberate how best to revise or replace this law, it is important that they consider the implications of a nationwide mandate of an educational practice, such as testing. Any educational practice, including different types of testing, originates from a vision of society and education. Each particular vision is grounded in certain assumptions, beliefs, and values orientations that reflect what the segment of the population that holds this vision believes to be the desired social and educational outcomes and preferred types of educational experiences for school children. Such practices are developed or selected because they align with this belief system, and because they are believed to contribute toward achieving the desired social and educational outcomes, and to provide the preferred type of educational experience for students.

For the purposes of this dissertation, a belief is an opinion or conviction. An assumption is something that is taken for granted. A values orientation is defined as a set or framework of beliefs, or world view, which tend to be accepted and adopted as a whole. It is important to note that beliefs, assumptions, and values orientations are all based upon opinion, and not necessarily evidence from tested hypotheses. Because all three are based on opinions, convictions, or

"beliefs," unless stated otherwise, the words "belief" or "beliefs" will refer to "assumptions, beliefs, and values orientations."

In addition, "social outcomes" refers to a beliefs about the kind of society that is desired. An example might be a society in which each individual achieves success based primarily on his or her merit. Similarly, "educational outcomes" refers to beliefs about the purpose of public education. An example of an educational outcome is that future workers will have a skill set required by business. Finally, the phrase "educational experiences" refers to the kinds of experiences that are desired for students in schools. An example of a preferred educational experience might be that students learn by solving problems. For the sake of simplicity of language, these will be referred to as "outcomes and experiences," unless otherwise stated.

By mandating a practice as part of federal law, lawmakers effectively endorse the specific set of beliefs from which the practice originated. In addition, they may be moving toward institutionalizing the corresponding outcomes and experiences associated with that practice. Therefore, lawmakers need to be aware of the beliefs that drive the practices that they incorporate into federal or state law. Furthermore, they need to be aware of the outcomes and experiences that practices such as testing are designed to advance. In this way, they will be able to pass education laws that support the beliefs, outcomes, and experiences that they intend to promote.

This dissertation focuses on the practice of mandated accountability testing. Its purpose is to explore and illuminate the beliefs that led to accountability testing proposals, as well as the outcomes and experiences that accountability testing proposals were intended to advance. This can help lawmakers to craft policies that align with or promote desired beliefs, outcomes and experiences for future generations of school children.

One way to explore and illuminate the beliefs, and desired outcomes and experiences that drive accountability testing is to analyze the arguments from both sides of the contemporary accountability testing policy debates leading to the passage of NCLB (2002). Such analysis of the arguments could aid in the discovery of these beliefs, outcomes and experiences that were expressed through these arguments. In addition, by comparing these arguments with those from similar policy debates involving different types of testing, it is possible to identify themes that were held in common between different types of testing, or which were unique to a specific type of testing. Such a comparative analysis between two similar situations in time is a historical/ comparative analysis (Babbie, 2004).

Historical/ comparative analysis has a long history in the social sciences (Babbie, 2004). Max Weber (1905/ 1958) used this method to develop his theory on the protestant ethic and the rise of capitalism. In addition, Karl Marx used a historical comparative/ analysis in developing his theory that economic systems influenced every other aspect of society (1843/ 1970). More recently, George Fredrickson (1981) used historical/ comparative analysis to compare apartheid in South Africa to slavery in the Southern United States. Similarly, David Tyack and Larry Cuban (1995) used historical/ comparative analysis to gain insights into the expectations that the public tends to place on schools to solve social problems, and the incremental process by which school reform takes place in the United States.

An apposite historical comparison to the contemporary accountability testing debates is the series of debates over the appropriate use of intelligence tests. The intelligence testing debates occurred during the early 1920s (Block & Dworkin, 1976; Cravens, 1986; Cremin, 1961/ 1964; Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Sacks, 1999; Sokal, 1987/ 1990; Thomas, 1982), with resurgences in 1969 (Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006;

Jacoby & Glauberman, ed. 1995; Montagu, 1975/ 1999a), and again in 1994 (Gould, 1981/ 1996; Gresson, 1996/1997; Jacoby & Glauberman, ed. 1995; Jackson & Weidman, 2004/ 2006; Montagu, 1997). The intelligence testing debates focused on the possible benefits, costs, and unintended consequences of using intelligence tests to classify and sort students according to what was believed to be innate ability, and then use the results to place students into educational tracks.

The intelligence tests were believed to measure ability or aptitude. They were thought to predict potential. By contrast, the achievement tests used for accountability testing were believed to measure what students had learned so as to measure the effectiveness of schools in teaching students. Comparisons of the beliefs expressed through the arguments from the respective testing debates should provide insights into themes common to all types of testing, as well as themes unique to a specific type of testing. These themes may provide insights into the beliefs, outcomes, and experiences valued by those who embraced testing as an educational reform compared to those who were critical of testing. Such insights can help lawmakers to align future testing policies and education laws in ways that promote positive outcomes and experiences for students.

Method

This study is a historical/ comparative analysis comparing the arguments from the accountability testing debates of the 1980s, 1990s, and 2000s to the arguments from the intelligence testing debates of the 1920s, 1970s and 1990s. Arguments for and against testing in the respective debates will be analyzed and compared in order to identify emerging themes common to the different testing policy proposals. The analysis will adhere to the following

structure. Chapter 2 provides a historical overview and contextual information of the events leading to the respective testing movements. Chapter 3 is an analysis of the arguments from the intelligence testing debates of the 1920s. Chapter 4 provides an analysis of the arguments from the more recent intelligence testing debates that resulted from publications of *How Much Can We Boost IQ and Scholastic Achievement* (1969) and *The Bell Curve: Intelligence and Class Structure in American Life* (1994), respectively. Chapter 5 is an analysis of the arguments of the criticisms of public education and the testing for accountability debates that followed. Finally, Chapter 6 is an analysis of the themes that emerged from the debates, and their implications for future testing policy.

Several interrelated themes emerged from the analysis that follows. These themes include: 1) merit; 2) "race," class and educational equity; 3) the meaning of democracy; 4) the fundamental purpose of public education and desired educative experiences in the United States; 5) the role of science and ideology in policy making; and 6) the tendency to oversimplify.

Sampling and Historical Interpretations

In conducting this study, it was necessary to select a manageable sample of published works from the participants in the different debates. As Gaddis (2002) wrote, one way to look at history is as a continuous and seamless series of an infinite number of concurrent particular events. It is impossible for the historian to describe all of the events associated with a historical period. As such, the historian's challenge is to choose those specific events that best tell the story or support the historian's argument or interpretation of the past. These arguments and interpretations can sometimes be generalized so as to apply to similar conditions in the past, present or future. Different historians can use different specific events to arrive at different

interpretations of the same time period. Indeed, George Herbert Meade (1964) argued that history is invented, and constantly reinvented based upon new information, as well as present and future needs. By choosing particulars from the past to include for study, the historian creates an interpretation of the past (a history) that reflects the present context as well as the historian's purpose and viewpoint.

Therefore, this analysis must be limited in scope to a manageable size, while still including a representative sample of those individuals who figured prominently in the respective debates. As such, this analysis will focus primarily only on the published works of participants in the respective testing debates. These works were selected based upon an initial review of literature, as well as the recommendations of professors at the University of Kansas. In addition a type of snowball sampling technique was employed, wherein references of selected works led to the discovery of other participants in the debates (Babbie, 2004; Merriam, 2009). Finally, works from the participants of these debates were reviewed. Those works that did not directly address issues related to the research questions were excluded from the analysis.

Using the methods described above, works were sampled from the following proponents of intelligence testing from the 1920s: Lewis Terman (1919, 1920, 1922a, 1922b), Henry Herbert Goddard (1922), Guy M. Whipple (1922), and Carl Brigham (1923). Critics of intelligence testing included in the analysis were William Bagley (1922a, 1922b, 1925), Walter Lippmann (1922a, 1922b, 1922c, 1922d, 1922e, 1922f, 1923a, 1923b, 1923c), John Dewey (1922/ 2008a, 1922/ 2008b), and Horace Mann Bond (1924a, 1924b).

Although the debates subsided after 1930, they experienced two resurgences in the latter half of the Twentieth Century: once in 1969 with the publication of *How Much Can We Boost IQ and Scholastic Achievement* in the Harvard Educational Review (1969); and again in 1994 with

Richard Herrnstein and Charles Murray's publication of *The Bell Curve: Intelligence and Class Structure in American Life*. The authors of both works took the position that intelligence was innate and relatively immune to environmental experiences. Both argued that the observed racial differences in average intelligence test scores were caused by genetic factors endemic to the particular "race." And both argued that attempts at melioration through compensatory education would have little lasting effect, and were therefore a waste of resources.

Although 25 years separated these two publications, the arguments put forward by Jensen (1969) and Herrnstein and Murray (1994) were very similar, as were the subsequent critical responses. Therefore, the debates that sprang from these publications will be analyzed together. The proponents of intelligence testing during these renewed debates will include Arthur Jensen (1969), Richard Herrnstein and Charles Murray (Herrnstein & Murray, 1994). Critics will include the following: John Rury (1995); Stephen Jay Gould (1996/ 1981, 1999); James Heckman (1995); Leon Kamin (1999); Charles Lane (1999/1975); Alan Ryan (1999); Richard Lewontin (1970/ 1975/ 1999); Peggy Sanday (1999/ 1975/ 1972); S. Biesheuvel (1999/ 1975/ 1972); Urie Bronfenbrenner (1999/ 1975/ 1972); Ashley Montagu (1975/ 1999a, 1975/ 1999b, 1975/ 1999c, 1997); Edmund Gordon and Derek Green (1999/ 1975/ 1974); C. Loring Brace and Frank Livingstone (1999/ 1975/ 1972); Leonard Lieberman, Alice Littlefield and Larry Reynolds 1999); W.F. Bodmer (1999/ 1975/ 1972); Jerome Kagan (1999/ 1975); S.E. Luria (1974/ 1975/ 1999); Steven Barnett (1995); and Lee, Brooks-Gunn, Schnur, and Liaw (1990).

Because the critique of public education led to calls for standards, testing and accountability, and because several of the critics were also promoters of standards, testing and accountability, critics of public education during the 1980s and 1990s will be considered alongside the promoters of testing for accountability. This group includes David Kearns (1988;

Kearns & Doyle, 1989); Chester Finn (Finn, 2002; Finn & Ravitch, 1996; Finn, Kanstoroom, Rothstein, & Honig, 2001; Finn, Manno, & Vanourek, 2001; Kanstoroom & Finn, 1999); Diane Ravitch (1993, 1994, 1996; 1999; Finn & Ravitch, 1996); Eric Hanushek (1989, 1994, 2003 2005; Hanushek & Raymond, 2005); Herbert Walberg (1994, 1998); and Marshal Smith and Jennifer O'Day (1991). In addition, works which describe minority advocacy groups support for standards, testing, and accountability will be included to add the perspective of these groups (Archer, 2006; Reid, 2005; Rury, 2012; Salzman, 2006).

Finally, defenders of public education and critics of standards, testing and accountability include the following: Gerald Bracey (1987, 1996); Susan Ohanian (1999, 2000, 2003); Alfie Kohn (2000; 2004a, 2004b; 2004c); Richard Rothstein (Finn, Kanstoroom, Rothstein, & Honig, 2001; Rothstein, 2004, 2008; Rothstein, Jacobson & Wilder, 2008); Peter Sacks (1999); Linda McNeil (1988 a, 1988b, 2000); Linda Darling-Hammond (1990; 2000, 2004, 2006); Alfie Kohn, 2000, 2004a, 2004b); Daniel Koretz (2008); George Madaus (Madaus, 1988; Madaus, Russell & Higgins, 2009); Deborah Meier (2004); George Wood (2004); and Stan Karp (2004).

Limitations

While this analysis potentially offers valuable insights, some caution is required when interpreting the results. While an attempt was made to select a representative sample of writers on both sides of the respective testing debates, it is possible that one side or the other is more heavily represented, or that important participants in the debates were omitted. Furthermore, much of the writing analyzed was an expression of the opinion held by the individual writer that was sampled, rather than evidence from research. Where research was presented by these writers, their interpretations frequently reflected *a priori* held conclusions.

Finally, this analysis is itself an interpretation. While an effort has been made to provide a balanced treatment of the various arguments, it should be noted that all historical interpretations reflect the biases of the writer. For example, a number of themes emerged during the analyses in chapters 3, 4, and 5, respectively, which were excluded from the discussion in Chapter 6. Some examples of these themes are "empowerment and locus of control," "distortion," and "test validity," to name a few. Those themes that were included in Chapter 6 were selected because they were of interest and were interrelated, thereby providing a focus to the discussion and conclusions. Themes that were not interrelated are topics for subsequent research. Furthermore, some of the themes identified in the analyses from individual chapter could be combined with themes from other chapters. Finally, it is possible that other researchers would have identified different themes from the same selection of writing.

Summary

Educational practices flow from a set of beliefs. They are designed and promoted to advance certain desired outcomes and experiences. Federal education laws that mandate certain practices can have the effect of endorsing the associated values, or institutionalizing the practices that are designed to produce certain outcomes and experiences.

Therefore, it is important for policymakers to be aware of the beliefs and intended outcomes that lead to certain educational practices. In this way, they can craft educational laws that are aligned with a desired set of beliefs and outcomes.

The purpose of this dissertation is to illuminate the beliefs associated with accountability testing, as well as the outcomes and experiences that testing is intended to advance. This study uses comparative/ historical analysis to compare samples of arguments from the intelligence

testing debates with samples of arguments from the contemporary testing for accountability debates, in order to explore the beliefs, outcomes, and experiences that are expressed in the arguments of the respective debates.

The analyses that follow led to the identification of six interrelated enduring themes. These themes include: 1) merit; 2) "race," class and educational equity; 3) the meaning of democracy; 4) the fundamental purpose of public education and desired educational experiences in the United States; 5) the role of science and ideology in policy making; and 6) the tendency to oversimplify. These themes and their implications for policy makers will be discussed in the final chapter.

Chapter 2: Testing debates in context

Introduction

Both intelligence and accountability tests and their respective policy debates emerged from specific historical contexts. Contextual factors leading up to and concurrent with the emergence of a particular testing proposal helped to shape the beliefs of the individuals who promoted or criticized the proposed uses for the respective tests. Furthermore, contextual factors leading up to the emergence of testing proposals can help to explain how and why these proposals came about, as well as how these tests gained ascendancy, for a time, as a popular tool for reforming education. As such, an analysis of the historical context leading up to and concurrent with the emergence of the respective testing movements can provide important background information critical for understanding the analyses of the arguments from the respective testing debates that will follow. This chapter describes the historical context that led to testing proposals, and the corresponding debates.

The Influence of Evolution.

In 1859, Charles Darwin published *On the Origin of Species* (1859/ 1902), in which he argued that new genetic traits continuously and randomly emerged within organisms. Traits that offered a survival advantage were more likely to be passed on to offspring. Organisms with traits that caused a survival disadvantage were less likely to survive to reproduce, causing the trait to die out. This theory explained how a species changed over time and new species emerged.

The theory of evolution was very influential, not only in the natural sciences, but in the emerging social sciences, as well (Baker & Stites, 1991; Cravens, 1978/ 1988; Minton, 1987/ 1990; Sokal, 1987/ 1990a). While a misinterpretation of Darwin, there was an idea that

organisms “improved” through natural selection, and that the different species of organisms were the living record of evolution, from the lowest to the highest forms of life (Cravens, 1978/ 1988; Reed, 1987/ 1990, p.79). This idea was applied to human “races,” with different “races” considered to represent a living fossil record that could be ranked along an evolutionary scale (Reed, 1987/ 1990). This idea that organisms were improved over time was not only applied to human beings, but also to human institutions. Indeed, the term “Progressive” implied that human institutions were also “evolving” toward a more advanced state (Brinkley, 2004).

Minton argued that social thought during the progressive era was dominated by Darwin’s theory of evolution (Minton, 1987/ 1990). In particular, the influence of evolution helped to define the intelligence testing debates. Minton suggested that participants in these debates fell into one of two camps: Social Darwinism and Reform Darwinism.

Social Darwinism was a form of biological determinism. Social Darwinists believed that biological processes had a determinist effect on social phenomena. Lewis Terman, Henry Herbert Goddard, Guy Whipple, and other advocates of intelligence testing subscribed to the belief that “natural endowment” determined intelligence, and intelligence determined whether the individual achieved academic, economic, and social success. In addition, biologically produced intelligence determined one’s level of moral behavior. Social Darwinists believed that social interventions would do no good, and may actually have dysgenic effects by allowing defective humans to survive and reproduce (Minton, 1987/ 1990).

By contrast, Reform Darwinists believed that social forces could shape society. As such, social interventions could help society to progress toward a better state. Reform Darwinists included critics of intelligence testing, such as John Dewey and William Bagley (Minton, 1987/ 1990).

Mental Measurement and the Emergence of Eugenics

At least some of Darwin's influence was through his half cousin, Sir Francis Galton (Jackson & Weidman, 2004/ 2006; Murdoch, 2007; Gould, 1981/ 1996; Sacks, 1999; Sokal, 1987/ 1990b). Through his reading of *On the Origin of Species*, Galton formed the theory that inherited mental traits determined life success (Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Murdoch, 2007; Sokal, 1987/ 1990b). To test this theory, he identified the prominent men who were related to one another. While the numbers were relatively low, they were greater than would have occurred by chance. He also found that eminent men who were related to other eminent men tended to have gained eminence in the same area, leading him to conclude that ability was inherited. Galton published the results in *Hereditary Genius* (Galton, 1870; Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Murdoch, 2007).

Because of his belief that social rank, income, and education were the manifestations of inherited ability, Galton believed that marginalized groups, including Blacks and women, occupied those positions as a result of their innate inferiority. Furthermore, he believed that the human "race" could be "improved" through selective breeding, or "eugenics," wherein successful individuals would be encouraged to mate with one another, while the lower classes would be discouraged from reproducing (Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Minton, 1987/ 1990; Murdoch, 2007).

Beginning with Galton, a number of early social scientists attempted to develop a means of assessing the inherent intelligence of human beings. They believed that intelligence was that which separated lower animals from humans, and separated lower humans from higher (Reed, 1987/ 1990).

Early attempts by researchers such as Galton and James McKeen Cattell focused on anthropometric tests. Anthropometrics involved the comparative study of "simple measurements of bodily dimensions," and properties (Murdoch, 2007; Sokal 1987/1990, p. 29). These tests were based on the assumptions that humans perceived the outside world through their senses and that genetically superior individuals would be physically superior as well as mentally superior (Murdoch, 2007; Sokal, 1987/ 1990b).

Cattell admired Galton's work. Like Galton, Cattell set out to devise a series of easily measured tests, hoping to find one that measured mental ability. However, he missed Darwin's larger point that variations might serve a function (Sokal, 1987/ 1990b).

With no real theory to guide him, Cattell believed that if he collected enough data on a variety of traits, he could find something that was a measure of mental ability (Sokal, 1987/ 1990b). However, not being mathematically proficient himself, Cattell persuaded one of his graduate students, Clark Wissler, to study correlation coefficients under the anthropologist Franz Boaz and then use them to determine whether there was a relationship between the various anthropometric measures and academic success.

Wissler found virtually no correlations between each of the anthropometric measures and class standing (Murdoch, 2007; Sokal, 1987/ 1990b). However, he did find that grades in one class correlated strongly with grades in other classes (Sokal, 1987/ 1990b). These findings effectively ended the movement to use anthropometric tests as indirect measures of intelligence.

General Intelligence (g)

At around the same time that Cattell and Wissler were attempting to find correlations between their anthropometric data and class standing in school, Charles Spearman was

conducting similar tests on school children on the Island of Guernsey, where he was stationed during the second Boer war. Spearman found strong correlations between grades in different courses, with correlations strongest in those courses that were presumed to require more "thinking," (Spearman, 1904).

These findings led Spearman to postulate that there was a unitary intelligence factor, something he called general intelligence (*g*), which determined an individual's success in a variety of intellectual tasks (Gould, 1981/ 1996; Murdoch, 2007; Sacks, 1999; Spearman, 1904; von Mayerhauser, 1987/ 1990). Spearman postulated that if *g* was a single trait that manifested itself in general intellectual ability, then a test made up of an almost random variety of different activities would be the best way to isolate and measure *g*. Furthermore, he considered intelligence to be *the result of a single factor*, a genome.

The Binet-Simon Test

The same year that Spearman published his two factor theory of intelligence, the French Minister of Education commissioned Alfred Binet to develop a test that could identify children who were behind their age cohort peers and who could benefit from some form of "special" education (Baker & Stites, 1991; Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Murdoch, 2007; Zenderland, 1987/ 1990). Binet and Theodore Simon developed a series of activities that they believed would differentiate typical children from those who were "retarded" in their development. There was no real theory behind their choice of items other than to select activities that required an ability to reason, as opposed to learned skills or knowledge (Baker & Stites, 1991; Gould, 1981/ 1996; Madaus, Russell, & Higgins, 2009; Murdoch, 2007; Sacks, 1999; Zenderland, 1987/ 1990). The original test was first published in 1905. However, in 1908

the test items were assigned an age level meant to identify the youngest age at which a typical child could successfully complete the task. By using items that differentiated between age cohorts, they believed that they were developing a test that would be useful to teachers (Murdoch, 2007).

Binet did not believe that he was measuring a single factor that was "intelligence," (Baker & Stites, 1991; Gould, 1981/ 1996; Madaus, Russell, & Higgins, 2009; Murdoch, 2007). In fact, he did not believe that intelligence could be measured, having written in 1905 that, "The scale, properly speaking, does not permit the measure of the intelligence because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured," (Binet as quoted in Gould, 1981/ 1996, p. 386). Nor did Binet believe that intelligence was innate and immutable (Gould, 1981/ 1996; Murdoch, 2007). Indeed, he believed that an "innatist" interpretation of his test would lead teachers to the conclusion that certain children could not learn, an effect that was diametrically opposed to his purpose in creating the test (Gould, 1981/ 1996). Binet believed that children who were behind their age peers needed more intense work, and was afraid that teachers who believed that intelligence was innate would simply give up on these children. Unfortunately, the meaning of the test results were reinterpreted by the American psychologists who translated and revised the Binet-Simon test to be a measure of an inborn and immutable quality that predicted educational, economic, and life success (Gould, 1981/ 1996).

Testing the Feebleminded

In 1906, a little known psychologist named Henry Herbert Goddard operated a lab at the Training School for Feeble-Minded Girls and Boys in Vineland, New Jersey. Goddard needed an

instrument to assess the Vineland students by degree of mental disability. He was using anthropometric tests five years after Cattell and Wissler published their paper demonstrating no relationship between anthropometric measures and mental abilities, not because he was ignorant of their results, but because he had no better test with which to replace them (Sokal, 1987/ 1990b). In 1908, Goddard traveled to Europe hoping to find a more effective alternative to the anthropometric tests. There, he learned of the Binet-Simon test, and soon realized that it was the assessment for which he had been searching (Murdoch, 2007; Zenderland, 1987/ 1990).

On returning to the United States, Goddard translated the test into English, and used it to assess the students at Vineland. On giving the tests to children at Vineland he found that the tests corresponded to staff members' intuitive assessments of the intelligence of the children (Zenderland, 1987/ 1990). Goddard became a convert to the new tests.

The Stanford-Binet Revision

Goddard may have brought the Binet-Simon Intelligence Test to the United States, but Lewis Terman is credited with popularizing intelligence testing in this country (Cremin, 1961/ 1964; Gould, 1981/ 1996; Minton, 1987/ 1990). In 1916, Terman published the Stanford-Binet revision of the Binet-Simon Intelligence test (Terman, 1916).

Terman's revision of the Binet test was the first to include the Intelligence Quotient (IQ), which assigned a standardized number to each examinee's score (Minton, 1987/ 1990). Prior to the introduction of the IQ, Binet's test assigned a "mental age" score that was based upon the average number of questions that a certain percentage (usually 75%) of a sample from a specific population of children of a given age was able to answer correctly (Fass, 1980). Terman adopted Stern's suggestion to divide the individual's mental age by his or her chronological age, resulting

in a ratio (Cremin, 1961/ 1964; Fass, 1980; Minton, 1987/ 1990). An IQ of 1.0 was equal to the average of the individual's age cohort. A score above 1.0 meant that the child scored above the mean relative to the age cohort reference group, while a score below 1.0 meant that the child scored below the mean relative to this age cohort reference group (Minton, 1987/ 1990). Later, Terman placed IQ scores on a standardized norm referenced scale with a mean at each age level of 100 and a standard deviation of roughly 15 points (Gould, 1981/ 1996).

Because IQ was expressed as a number, but not as a percentile rank, it presented the illusion that an IQ score was an absolute score based upon the achievement of scientifically determined criteria rather than a score that was relative to a specific population in a specific time and place (Fass, 1980). Furthermore, Terman established norms for his test using middle class white children from Palo Alto, California, and assumed that he had the proper mix of questions when his reference group achieved scores that were evenly distributed along the normal curve (Cravens, 1978/ 1988). Consequently, the standard was established on a sample that was not representative of, and probably more privileged than the majority of the children who would eventually be tested (Fass, 1980).

The Army Psychological Testing Program

At the turn of the century, psychology was a relatively new field. The second generation American psychologists struggled to establish psychology as a legitimate science on a par with the natural sciences, thereby raising the level of prestige enjoyed by members of the profession (Gould, 1981/ 1996; von Mayrhauser, 1987/ 1990). To Robert Yerkes, Harvard professor and then president of the American Psychological Association (APA), America's entry into World War I (WWI) presented an opportunity to promote his science, his profession, and, quite

possibly, himself, all while helping America's war effort (Gould, 1981/ 1996; Murdoch, 2007; Reed, 1987/ 1990).

Yerkes believed that this opportunity lay in the newly developed intelligence tests. Intelligence tests could be included in the battery of medical tests that the army used to screen young recruits. The purpose was to sort recruits by ability in order to assign them to appropriate positions (Jackson & Weidman, 2004/ 2006; Tyack, 1974), as well as to identify those unfit for military service (Baker & Stites, 1991; Murdoch, 2007; Reed, 1987/ 1990; Tyack, 1974).

The Binet-type individual intelligence tests were ill-suited for the task of assessing 1.7 million men in a short period of time. In order to make the testing more efficient, the team of psychologists that Yerkes had assembled adapted a group intelligence test first developed by Arthur Otis, a graduate student of Terman's (von Mayerhauser, 1987/ 1990; Minton, 1987/ 1990; Samelson, 1987/ 1990). This test used the newly developed multiple choice format, the development of which was credited to Frederick Kelly (Samelson, 1987/ 1990).

Ultimately, the Army Intelligence tests consisted of three tests. Two of which were adaptations of Otis's group intelligence tests, and a Binet-type individual test. The Alpha was for literate English speaking recruits. The Beta was for those who failed the alpha test, were illiterate, or were non-English speakers (Brigham, 1923; Gould, 1981/ 1996; Murdoch, 2007). Finally, a Binet-type individual test was chosen for retesting recruits who scored poorly on the Beta test (Brigham, 1923; Gould, 1981/ 1996).

The tests were validated by correlating the scores of a sample of 4000 men with ratings of their "soldierly value" as provided by their commanding officers (Murdoch, 2007; Tyack, 1974). Yerkes argued that if the tests correlated so well with officer opinion of soldierly value, then intelligence must be an important component for classifying soldiers (Murdoch, 2007). In

addition, the tests correlated with other tests of intelligence (Brigham, 1923; Goddard, 1922). The army agreed to allow Yerkes's group to test all recruits (Gould, 1981/ 1996; Murdoch, 2007).

By the end of the war, the psychologists had tested over 1.7 million men (Baker & Stites, 1991; Gould, 1981/ 1996; Murdoch, 2007; Tyack, 1974). Tyack wrote that because the tests were eventually used to track men into different military jobs, they potentially determined whether a man was given an office job in Washington, or was sent to the front lines (Tyack, 1974). However, others have argued that the recommendations of the psychologists were largely ignored (Gould, 1981/ 1996; Murdoch, 2007; Reed, 1987/ 1990; von Mayerhauser, 1987/ 1990).

Indeed, there is evidence that the army viewed the psychologists as more of a nuisance than a help (Gould, 1981/ 1996; Murdoch, 2007; von Mayrhauser, 1987/ 1990). Rather than giving psychologists the prestige of assisting doctors in the medical corps, as Yerkes sought, the psychologists were relegated to the ignoble Sanitary Corps (Murdoch, 2007; Reed, 1987/ 1990; von Mayrhauser, 1987/ 1990). While psychologists were commissioned officers, the army consistently commissioned them below other professionals (Murdoch, 2007). Psychologists were typically commissioned as lieutenants, while doctors were commissioned as "captains, majors, or lieutenant colonels." Finally, only between one and ten percent of soldiers rejected or discharged because they were found to be mentally unfit were rejected because of their scores on the Army intelligence tests. This suggested that the Army did not take the tests or the testers seriously (Samelson, 1977).

The Army tests themselves, and the protocol followed by the testers probably had an impact on the results (Gould, 1981/ 1996; Lippmann, 1922b; Murdoch, 2007). The tests were culturally biased, and many of the test items measured experience rather than innate ability

(Gould, 1981/ 1996; Murdoch, 2007). Furthermore, protocol required the testers giving the Beta test to use pantomime when testing illiterate recruits, regardless of whether the recruits spoke English, a practice that was likely confusing to the recruits (Gould, 1981/ 1996; Murdoch, 2007).

Even with their flaws, the standard protocols developed for conducting the three tests were frequently not followed (Gould, 1981/ 1996; Reed, 1987/ 1990). Because of very long lines for the Beta tests, standards were lowered for taking the Alpha test, resulting in many illiterates taking the written alpha exam. The psychologists also consistently failed to retest men who scored 0 points on one or more of the tests (Reed, 1987/ 1990). African American recruits were frequently not given the opportunity to take the Alpha test (Murdoch, 2007). African American recruits who did take the Alphas and who scored poorly were typically not allowed to take the Beta, despite the fact that those who did take both tests usually improved their scores on the Beta (Gould, 1981/ 1996). Similarly, Blacks who scored a 0 on the Betas were frequently not retested individually (Murdoch, 2007). Foreign born recruits were often not given the opportunity to take the Beta tests, despite their lack of proficiency with English (Murdoch, 2007). The inclusion of these scores in the results tended to bolster nativist and racist interpretations of the data (Reed, 1987/ 1990).

Other practices further biased the results of the Army testing program. Army psychologists were concerned about the large number of 0 scores on the alpha test items (Yerkes, 1921, p. 622). Indeed, many of the Alpha test items had a mode of 0. This suggested that large numbers of men did not understand what these items was asking them to do. However, rather than throwing these items out, Army psychologists treated the 0 score as though it were merely the lowest score that a man could possibly get on that item. Scores “piled up,” (Yerkes, 1921, p. 622) at 0 because the instrument could not measure lower scores. Psychologists assumed that in

reality, some men actually received bonus points that were in direct proportion "with his stupidity," (Yerekes, 1921, p. 622). Therefore, they chose to "calibrate" the man's score against his score on the other items in the series, by using linear equations (Yerkes, 1921, p. 623). If the man had done well on the other items, he kept his 0 score. If he had done poorly, his score on the item was converted to a negative number. Needless to say, this had a negative impact on aggregate scores.

The Army tests yielded tremendous amounts of data. These data were usually interpreted in ways that fit with the researcher's paradigm, even when other interpretations might seem to make more sense. For example, when Yerkes found that there was a high correlation between years of schooling and scores on the tests, he attributed these higher scores to the fact more intelligent people remained in school longer, rather than admit that the tests might actually measure educational experience (Cravens, 1978/ 1988; Gould, 1981/ 1996, 1991; Reed, 1987/ 1990).

He also found that recruits who did not speak English performed worse on the Alpha than on the Beta test, and that there was a correlation between the number of years that immigrants had been in the United States and their scores on the tests. He interpreted this to mean that those immigrants who were more intelligent tended to be more successful in the United States, and were therefore less likely to return to their countries of origin. Others attributed this finding to the superiority of Nordic "races" and the inferiority of Southern and Eastern Europeans, noting that earlier waves of immigrants were from Nordic countries, while later waves of immigration tended to be from Southern and Eastern Europe (Gould, 1981/ 1996).

What would have been a comical interpretation of the results had not so many people taken it seriously was the finding that the average mental age of the average army recruit was

three years below average! Rather than interpret this to mean that there was something wrong with the test, the norms, the norm reference group, or the testing protocol, Yerkes was disturbed by the prospect that stupidity was much more common than previously thought, and Goddard warned that this meant trouble for the democracy (Gould, 1981/ 1996).

Finally, what probably was the most tragic of the interpretations, and unfortunately most enduring, was that racial and ethnic groups could be ranked by the average group scores on the tests (Brigham, 1923). Brigham ranked groups by country of origin, "race," and average mental age as follows: England 14.87; Scotland 14.34; Holland 14.32; Germany 13.88; U.S. (White) 13.77; Canada 13.66; Sweden 13.30; Norway 12.98; Belgium 12.79; Ireland 12.32; Austria 12.27; Turkey 12.02; Greece 11.90; Russia 11.34; Italy 11.01; Poland 10.74; U.S. (Black) 10.41 (Brigham, 1923, p. 124). These group rankings were used to justify popular stereotypes and seemed to support the myth that the social hierarchy was based upon merit rather than "race" and ethnicity (Tyack, 1974).

Furthermore, because the majority of the psychologists who were involved in the army testing program were committed to a hereditary view of intelligence, they did not question these findings. Yerkes, for example, was committed to an evolutionary view of intelligence, in which organisms from the lowest form of life to human beings fell along an evolutionary continuum, and where different human "races" were at different points along this evolutionary continuum (Reed, 1987/ 1990). They did not consider the possibility that schooling, culture, English language proficiency, poverty, health and nutrition, or their own test and protocols may have contributed to these results.

Despite the many problems with the Army IQ testing program, these tests left a lasting legacy. Thanks to the skillful self-promotion of Yerkes and Terman, the army tests helped to

establish the legitimacy of the IQ and IQ testing (Gould, 1981/ 1996; Murdoch, 2007). Indeed, Reed stated that Yerkes's most "remarkable achievement was the myth that the army testing program had been a great practical success and that it provided a 'goldmine' of data on the heritability of intelligence," (Reed, 1987/ 1990, p. 84).

The army tests also marked the first time that tests were used on "normal" people to make decisions about them based upon ability (Reed, 1987/ 1990). The army testers incorporated new testing technologies into their tests that allowed for group IQ testing. This eventually led to the establishment of the United States testing industry (Gould, 1981/ 1996; Lemann, 1999/ 2000; Murdoch, 2007). Indeed, the National Intelligence Test (Fass, 1980; Murdoch, 2007; Tyack, 1974) and the Scholastic Aptitude Test (Lemann, 1999/ 2000; Murdoch, 2007) were both adaptations of the Army Alphas.

In addition, while the Army test results did not lead to the passage of the 1924 Immigration Restriction Act, they were used to help justify the passage of the law that assigned immigration quotas based upon country of origin (Gould, 1981/ 1996; Sokal, 1987/ 1990a). This act limited immigration to not more than 2% of the proportional racial and ethnic makeup of the United States population in 1890. This year was chosen because following 1890, there was a dramatic increase in the number of immigrants from Southern and Eastern Europe. Gould (1981/ 1996) argued that these quotas prevented many refugees who anticipated the Nazi Holocaust from immigrating to the United States prior to World War II (WWII).

The Cardinal Principles

If the WWI Army Intelligence tests provided the means to conduct mass intelligence testing on school children, the publication of *The Cardinal Principles of Secondary Education*

(The Commission on the Reorganization of Secondary Education of the National Education Association, 1918) provided the justification. While the *Cardinal Principles* never explicitly discussed testing, it set the stage for the testing and tracking movement by critiquing American education, identifying objectives for reform of education, translating a "social efficiency" mode of progressivism into a means for the reformation of education, and, most importantly, recommending vocational and college bound educational curriculum tracks based on ability and interests. The new group intelligence tests offered a means to rank and sort students by ability, thereby offering schools an efficient way to place students into vocational or college bound curricular tracks.

The "Social Efficiency" model of progressivism mentioned above was the belief that society could be made more efficient through the use of "scientific management," (Callahan, 1962). Scientific management involved the elimination of wasted time, effort, motion, and resources, in order to operate social entities more productively and with fewer resources. Social efficiency was related to scientific management in that it promoted the identification of the niche for which each individual was best suited, preparing that individual to fill that niche, and placing the individual in the niche. Callahan (1962) described how scientific management and social efficiency were incorporated into public education during the early years of the 20th Century, in order to make schools more efficient and productive.

Intelligence Testing in Schools

As early as 1908, researchers such as Thorndike (1908) and Ayers (1909) began calling attention to the large numbers of students who failed to advance through the grades at an acceptable rate. Ayers attributed this high failure rate to a variety of causes, including illness, late

start, poor attendance, or inadequate or poor enforcement of compulsory attendance laws. At the same time, urban schools were struggling to assimilate large numbers of heterogeneous schoolchildren entering school as a result of compulsory education laws. School systems were struggling to meet the increasing costs of educating children. As such, providing classroom space for students who were repeating a grade was seen as an inefficient use of resources. In addition, schools were facing increasing criticism for a lock step curriculum that overwhelmed some students, but kept others from reaching their full potential (Tyack, 1974).

Intelligence tests offered a solution to the problems faced by urban schools in the early 1920s (Tyack, 1974). Tests could be used to sort students into homogeneous classes (Haggerty, Terman, Thorndike, Whipple, and Yerkes, 1920; Tyack 1974). Such homogeneous classes would allow teachers to adapt the level of rigor of their curriculum and instruction to the ability of the students in each class. This would make education more efficient. In addition, tests could be used to provide vocational guidance to students, to identify especially gifted or dull students, and to diagnose students with learning problems.

The Army Intelligence Testing Program had demonstrated for the first time that intelligence tests could be administered to large numbers of individuals simultaneously (Tyack, 1974). It also represented the first time that intelligence tests had been used on "normal" populations, as opposed to the diagnosis of feeblemindedness. Therefore, the tests promised to offer a viable tool for classifying students into homogeneous classes, based upon ability.

In 1919, a number of psychologists from the Army testing program gathered to convert the Army Alpha into a group intelligence test designed for school children. The "National Intelligence Test," as it came to be called, was published in 1920 (Haggerty, et al., 1920).

Within six-months of the publication, the National Intelligence Tests sold 400,000 copies. A year later, Terman stated that over two million children had been tested on one of over a dozen tests that were on the market. By the mid-1920s there were over 75 published intelligence tests (Murdoch, 2001). By 1925, 215 cities used the tests to classify and sort students (Tyack, 1974).

Tyack (1974) suggested that it did not matter whether or not the tests were valid measures of intelligence. The tests provided an efficient tool to aid in the administration of a school system In addition, the numeric score gave the appearance of objectivity, making it easier for teachers and administrators to convince parents to accept decisions that were made based upon the scores.

The IQ Testing Debates

The publication of the Army test results by Yerkes (1921) and Carl Brigham (1923) along with the publication of *The Cardinal Principles* (The Commission on the Reorganization of Secondary Education of the National Education Association, 1918) sparked a series of debates over the proposed use and interpretation of results of the new tests (Block & Dworkin, 1976; Cravens, 1978/ 1988; Cremin, 1961/ 1964; Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Minton, 1987/ 1990; Sacks, 1999; Sokal, 1987/ 1990a; Thomas, 1982). Prominent among these debates were those between William Bagley (1922a, 1922b), Lewis Terman (1922a), and Guy Whipple in 1922, and those between Walter Lippmann (1922a, 1922b, 1922c, 1922d, 1922e, 1922f, 1923a), John Dewey (1922/ 2008a, 1922/ 2008b), and Lewis Terman (1922b, 1923) in 1922 and early 1923. In addition, Horace Mann Bond (1924a, 1924b) and William Bagley (1925) responded to Carl Brigham's interpretation of the World War I Army Testing program results (Jackson & Weidman, 2004/ 2006; Thomas, 1982). The debates centered on the

heritability of intelligence, racial interpretations of average group IQ test scores, the validity of the tests, the nature of democracy and meritocracy, whether IQ tests actually measure intelligence, indeed, the very definition and nature of intelligence (Minton, 1987/ 1990).

In addition to these earlier debates, the intelligence testing controversy experienced occasional resurgences, most notably those sparked by the 1969 publication of Arthur Jensen's controversial essay, *How much can we I.Q. and scholastic achievement* (Baker & Stites, 1991; 1969; Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Jacoby & Glauberman, 1995; Montagu, 1975/ 1999a), and again in 1994 with the publication of Herrnstein and Murray's *The Bell Curve: Intelligence and Class Structure in American Life* (Gould, 1981/ 1996; Gresson, 1996/ 1997; Jackson & Weidman, 2004/ 2006; Jacoby & Glauberman, 1995; Montagu, 1997). The arguments from these debates will be analyzed in the following chapters. Before turning to the analysis of the IQ testing debates, however, it is also necessary to first set the context in which the standards, testing and accountability debates occurred.

Testing for Accountability

The idea of using student achievement test scores as a measure of school quality is not a new one. As early as 1895, muckraking journalist and pediatrician Dr. Joseph Mayer Rice used student performance on spelling tests to compare the effectiveness of instruction at different schools across the United States (Cremin, 1961; Haertel & Herman, 2005). Not long after, New York began using achievement tests to assess education in its public schools (Haertel & Herman, 2005). Other cities soon followed.

Around 1914 to 1915, Frederick Kelly is credited with having developed the first multiple-choice test (Samelson, 1987/ 1990). Arthur Otis adopted this innovation to create the

first group intelligence test, a variation of which was used during the World War I Army Testing Program. This development, along with the invention of the grading machine, made it possible to efficiently test large numbers of students (Lemann, 1999/ 2000).

The first state-wide assessment program was initiated by E. F. Lindquist in 1929, when he helped launch the Iowa Test of Basic Skills (Haertel & Herman, 2005; Lemann, 1999/ 2000). The original purpose of this test, however, was less one of assessing school quality than it was to identify students who could benefit from educational intervention, and thereby extend education to more children (Lemann, 1999/ 2000).

Prior to the 1950's, most decisions impacting public education were made at the state or local level. This tradition of local control began to change with the *Brown v. Topeka Board of Education* decision of 1954 (*Brown v. Board of Educ.*, 1954), and the passage ESEA (1965) (Baker & Stites, 1991; McGuinn, 2005).

The *Brown* (*Brown v. Board of Educ.*, 1954) decision forced increased federal oversight in education with the ruling that segregation of children by "race" in public schools violated the due process clause of the Fourteenth Amendment of the Constitution of the United States. Furthermore, as courts around the country began requiring school districts to racially integrate school systems, it became necessary for the federal government to enforce these decisions (McGuinn, 2005).

The Civil Rights Movement and the Johnson Administration's "War on Poverty" also exposed educational inequities related to "race" and socio-economic status. In general, states have historically funded schools using taxes on property. As a result, property rich school districts tended to have more resources to spend for education than property poor school districts. Consequently, there were large inequities between different school districts in the resources

available for schools (Baker & Stites, 1991; Berliner & Biddle, 1995; McGuinn, 2005). Furthermore, poor communities and poor families tended to have fewer resources available to help insure that children were school ready. These resources included such basics as healthcare and proper nutrition, in addition to quality daycare that provided enriching environments for young children (Berliner & Biddle, 1995; Rothstein, 2004). Such inequities in educational and community services provided to poor and minority children led to the passage of ESEA (1965) (Baker & Stites, 1991; McGuinn, 2005).

The ESEA (1965) was notable in that it was the first "major involvement of the federal government in funding and directing public education," (Koretz, 2008, p. 55), although the law was limited in scope to improving educational opportunity for poor schools and poor children (McGuinn, 2005). The law was based on the assumption that while most American schools were doing well, some local governments were either unable or unwilling to provide equitable resources to schools serving poor and minority communities (McGuinn, 2005). Thus, ESEA (1965) continued the tradition of local control for most schools, but allowed the federal government to intervene where local governments failed.

ESEA (1965) was also important because it marked the first time that the federal government mandated testing to evaluate service delivery (Baker & Stites, 1991; Haertel & Herman, 2005; Koretz, 2008). With the 1974 passage of the Title I Evaluation and Reporting System (TIERS, 1974), schools were required to use student scores on standardized, norm referenced tests to evaluate Title I programs (Koretz, 2008).

Focus on educational equity was tempered by the release of the *Equality of Opportunity Report*, popularly known as *The Coleman Report* (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966). This report found that schools and resources provided to

schools had little impact on student achievement that was independent of the child's home environment and social context (Baker & Stites, 1991; Coleman, et al., 1966; Haertel & Herman, 2005).

Finally, federal involvement in assessing schools was solidified with the introduction of the National Assessment of Educational Progress (NAEP) in 1969 (Rothstein, Jacobson & Wilder, 2008). Sometimes called "The Nation's Report Card" the original NAEP was purposefully designed to provide a snapshot of education at the national level without providing information at the state, district, school, or student level (Hazlett, 1974). This was intended to preserve local control over the curriculum. In addition, the original NAEP could assess any educational goal to which a school devoted 15-20 percent of its time (Hazlett, 1974). This was intended to prevent the NAEP from inadvertently narrowing the curriculum.

Civil Rights and Ability Testing

The Civil Rights Movement did not just focus attention on the inequitable distribution of educational resources. It also led many Americans to voice concerns over the disparate impact that some high stakes tests had on certain minority groups. Charges of cultural bias in the IQ and achievement tests then being used to make high stakes decisions for students resulted in a number of calls for testing reforms, as well as calls for a moratorium on certain uses of testing. In 1968, the National Black Psychologists called for high stakes standardized tests to be revised so as to be culturally fair (Baker & Stites, 1991). This was followed by the National Education Association's (NEA) call for a moratorium on standardized testing during the 1972/ 1973 school year. In both 1972 and again in 1974, the National Association for the Advancement of Colored People (NAACP) called for an end to the use of any tests that had not been corrected for cultural

bias (Baker & Stites, 1991). In 1975, the National Association of Elementary School Principals (NAESP) joined the critiques of standardized testing (Baker & Stites, 1991). Finally, in 1980, Ralph Nader's consumer protection group published a report critical of Educational Testing Services (ETS) that led to New York's *Truth in Testing Law* (Baker & Stites, 1991). This law required test manufacturers to publish the answers to annual test questions following test administration, so that consumers could evaluate test items for fairness.

Minimum Competency Exams

During the early 1970s, reports of declining test scores and high school graduates who were unemployed raised concerns that students were simply being passed along or allowed to graduate regardless of their achievement. Such reports raised concerns that a high school diploma was meaningless. They also began a shift in the focus of educational reform from inputs to outcomes (Haertel & Herman, 2005). As a result, beginning in 1971 states began passing minimum competency laws (Koretz, 2008). These laws required students to score above a minimum cut score in order to either be promoted to the next grade level or to graduate from high school. Although the tests tended to be relatively easy, they brought serious consequences for students who failed (Koretz, 2008). By 1987, forty states had minimum competency laws. However, Baker and Stites (1991) argued that a number of court rulings prompted states to lower the rigor of most of these exams to the point where they became essentially meaningless.

Minimum Competency exams were significant to the testing and accountability movement for several reasons. First, they initiated a growth in state testing programs that continued after minimum competency exams were discontinued. Second, they promoted a growth in criterion referenced tests. Finally, minimum competency exams marked a shift from

using tests to provide educators with information about students, to one of holding students and educators accountable (Koretz, 2008).

Conservative Voices

For 36 years following the passage of the 1965 ESEA, democrats wanted to limit the Federal role in education to one of increasing resources, or inputs, rather than focusing on improving outcomes for students (Debray, 2005; McGuinn, 2005). Conversely, conservatives were more concerned with States' rights, and sought to protect local control of education by eliminating federal intrusion in the form of excessive regulations. McGuinn (2005) argued that Republicans favored standards, testing and accountability, but believed that these and other educational policies should be left to the states.

Throughout the 1970s and 1980s, the Democrats continued to gradually expand the role of the federal government into educational policy. However, Democrats limited this expansion to increased resources tied to compliance with procedures and regulations aimed at improving educational equity (McGuinn, 2005). Conversely, Republicans became firmer in their opposition to federal involvement in education, and were increasingly critical of many public school practices (Debray, 2005; McGuinn, 2005).

Berliner and Biddle (1995) suggested that criticism of public education escalated during the 1980s and 1990s as conservatives gained influence and power, and that these attacks were motivated, at least in part, by a desire to promote the respective agendas of different conservative factions. Berliner and Biddle called this period of heightened criticism of American public education "The Manufactured Crisis," suggesting that the crisis in education was created to serve

political ends. This period of increased criticism of public education was initiated by the release of *A Nation at Risk* (Commission for Excellence in Education, 1983).

A Nation at Risk

A Nation at Risk (Commission for Excellence in Education, 1983) was released early in the Reagan administration. It used inflammatory language and military comparisons to create a sense of urgency over the state of public education in the United States, and blamed America's lost dominance in the world economy to the decline of American education.

A Nation at Risk (Commission for Excellence in Education, 1983) provided twelve indicators demonstrating the decline of the nation's public school system. These indicators included the following: 1) students in the United States ranked poorly compared to students from other countries on tests of international comparisons; 2) 23 million American adults were functionally illiterate; 3) 13 percent of American 17 year olds were functionally illiterate, with illiteracy rates among minorities in this age group as high as 40 percent; 4) the average score for high school students on standardized tests in 1983 was less than it was in 1957; 5) the achievement of over half the gifted population did not match their tested ability; 6) between 1963 and 1980, the average Scholastic Achievement Test (SAT) verbal score fell by 50 percent, and the average math score fell by 40 percent; 7) College Board achievement tests documented a consistent decline in physics and English during recent years; 8) the absolute number and proportion of students scoring 650 or higher on the SAT declined dramatically (no magnitude or time period given); 9) 17 year olds lacked "higher order thinking skills;" 10) Between 1975 and 1980, remedial math courses at public universities increased 72 percent; 11) average tested achievement of college graduates had declined; 12) business and military leaders reported that

they spent millions on remediation programs for new hires/ recruits. In addition, *A Nation at Risk* (1983) suggested that the need for workers with expertise in the fields of science and technology was increasing dramatically.

Based on these indicators, *A Nation at Risk* (Commission for Excellence in Education, 1983) made a number of recommendations. These recommendations called for more rigorous academic and behavioral standards; a focus on basic skills, emphasizing "The Five New Basics" of English, mathematics, science, social Studies, and computer science; more rigorous standards for entering the teaching profession; and the adoption of state tests designed to hold students accountable for their learning.

The recommendations also called for increased accountability at all levels of education, but fell short of recommending that educators be held accountable for student test scores. Finally, while these recommendations were a move toward a stronger role of the Federal government, they stopped short of calling for national standards or a national test.

Diane Ravitch (2010) defended *A Nation at Risk* (Commission for Excellence in Education, 1983), arguing that its blunt presentation of facts concerning the deterioration of the nation's school system became controversial because it raised the level of concern of the nation, which naturally made some educators feel defensive. However, critics of *A Nation at Risk* (Commission for Excellence in Education, 1983) argued that the authors intentionally used inflammatory language and questionable evidence to promote a conservative educational agenda (Berliner & Biddle, 1995). Others have suggested that it merely represented the periodic reorientation of American education from a student centered orientation to a state centered orientation, and that this reorientation toward the state was necessary to pave the way for the professional empowerment of teachers (Darling-Hammond & Berry, 1988).

Systemic School Reform

A Nation at Risk initiated an era of education reform (Baker & Stites, 1991). Initially, these reforms were state centered and focused on establishing rigorous standards, strengthening assessments, and regulating teacher behavior. They also included a number of strategies intended to improve educator quality (Darling-Hammond & Berry, 1988). Finally, they included mechanisms to hold educators accountable for student test scores (Baker & Stites, 1991). These state centered reforms began to take shape in the late 1980s with the Governor's Educational Summit (Debray, 2005; Debray-Pellot & McGuinn, 2009).

In 1989, President George H.W. Bush convened a panel of the governors from all 50 states in Charlottesville, North Carolina for an education summit (Debray, 2005; Debray-Pellot & McGuinn, 2009). Led by then Arkansas governor William Jefferson Clinton, the education panel identified six broad goals, which, if pursued, would lead to the development of rigorous standards and assessments.

Beginning in the early 1990s, state centered educational reform began to take shape around systemic school reform (Smith & O'Day, 1991). Smith and O'Day (1991) argued that previous education reforms, with their focus on giving educators the freedom to make professional decisions while holding them accountable for the results, was not likely to work on a widespread basis. This was in part because the education system was so fragmented, and partly because educators had been trained in a "fact-based" philosophy of knowledge, "hierarchical approaches to skill development, and a near total reliance on teacher-initiated and teacher directed instruction," (Smith & O'Day, 1991, p. 234).

Instead, Smith and O'Day (1991) advocated for a system-wide reform effort, beginning at the state level. In their vision, states would establish curriculum frameworks, and develop

standardized tests that were aligned to these frameworks. These tests would hold educators accountable for student learning along the curricular frameworks. Accountability tests and the corresponding promise of rewards or sanctions would motivate educators at the state and local level to align all facets of the education system to the curriculum frameworks, including district curricula, teacher pre-service training, teacher professional development, and student learning materials. In this way, standards, testing and accountability would create system-wide reform of public education.

Bill Clinton and the New Democrats

As president, Bill Clinton continued to promote systemic school reform through standards and accountability as his administration's education reform policy. Following his election as president in 1992 and with the help of a Democrat controlled House and Senate, Clinton was able to push through two education reform laws during his first two years in office (Debray, 2005; McGuinn, 2005).

The first of these two laws was *Goals 2000: Educate America Act of 1994* [Goals 2000](1994). *Goals 2000* (1994) created a framework and funding for systemic school reform. The law encouraged states to create standards and benchmarks in core content areas, and to develop assessments that were aligned with those content standards and benchmarks. There was to be a process for instructing the teachers in the standards, and aligning local curricula and materials with the standards. Finally, states were also required to develop their own plans for improving schools that failed to meet content standards and timelines (Haertel & Herman, 2005).

The second law passed early in the first Clinton administration was the *Improving America's Schools Act of 1994* [IASA] (1994). This law was the reauthorization of the ESEA of

1965. IASA marked the first time that standards, assessments, adequate yearly progress, school report cards, and corrective action were codified into federal education law (Debray, 2005; Haertel & Herman, 2005; McGuinn, 2005). The law required states to use proficiency standards in describing how well children had learned the standards. The law codified the use of assessments to evaluate the local education authority and schools (IASA, 1994). This law also encouraged states to offer school choice through vouchers and the establishment of charter schools (McGuinn, 2005). Finally, the IASA (1994) required state tests to measure critical thinking skills, and to employ multiple measures to assess student learning (Haertel & Herman, 2005).

The two laws did not carry mandates for states. However, they represented a sea change in federal education policy in that they moved away from federal focus on ensuring equity for marginalized students to one of improving academic performance of all students and all schools (McGuinn, 2005). This cemented the shift from reform efforts that focused on inputs to reforms that focused on outcomes (Haertel & Herman, 2005). However, lingering Republican opposition to federal involvement and Democratic opposition to testing and accountability left both of the laws "relatively weak and weakly enforced" (McGuinn, 2005, p. 52).

Performance Assessments

The emphasis on multiple-choice testing during the 1980s led to a growing realization that the design of the test narrowed the learning experience of the students to experiences that mimicked the test questions (Haertel & Calfee, 1983). This realization that test design communicated learning goals led to a movement for performance assessments to be incorporated into high stakes tests. Performance assessments were designed to promote critical thinking, and

typically required students to create something, solve a problem, and involved real-world (authentic) applications. Test items typically took longer to complete than multiple-choice tests (Herman, Aschbacher & Winters, 1992; Wiggins, 1992). Ultimately, these test items proved to be less reliable than multiple-choice tests, and therefore required many more tasks to gain an accurate assessment of individual students (Baker, 1994; Dunbar, Koretz & Hoover, 1991; Shavelson, Baxter & Gao, 1993). These problems with reliability and the additional time required to conduct performance assessments contributed to their infrequent use in high stakes testing. The demands of additional testing that were placed on schools by NCLB (2002) further diminished the use of performance assessments (Haertel & Herman, 2005).

Republican Backlash, and the Backlash to the Republican Backlash

The passage of these two education bills, along with other "New Democrat" policies, contributed to the rise of a Republican backlash (Debray, 2005; McGuinn, 2005). Led by Newt Gingrich, the 1995 Republican revolution and the corresponding "Contract for America" emphasized devolution, calling for cuts to education spending, block grants, and the elimination of the Department of Education. These Republicans tried to convince the public that federal involvement in education was harmful to the public schools, and that increased funding for schools was not related to better student achievement. Furthermore, they increasingly called for vouchers and other market based educational reforms as answers to the inequitable education provided by America's public schools. While these ideas appealed to the Republican base, they were not as popular with the public, at large. As a result, a coalition of Democrats and moderate Republicans were able to block the passage of most of this agenda (McGuinn, 2005).

During the late 1990s, polls indicated that Republican ideas for education reform had little traction with the American public (Debray, 2005; McGuinn, 2005). At the same time, minority groups were pushing democrats for more meaningful reforms than simply providing funds without accountability. President Clinton expressed this sentiment when he wrote that "the fundamental lesson of the last seven years, it seems to me, is that education investment without accountability can be a real waste of money. But accountability without investment can be a real waste of effort. Neither will work without the other. If we want our students to learn more we should do both," (Clinton, 2000).

Republicans and Democrats Converge

President George W. Bush's experience as governor of Texas convinced him that the key to improving education was through standards and accountability. However, he had to reconcile this with the traditional Republican position of advocating for decentralization and local control (McGuinn, 2005). Bush's education bill, NCLB (2002) attempted to reconcile these competing tensions by mandating that states set their own standards, create their own tests, establish their own Annual Yearly Progress (AYP) goals, and set their own criteria for teachers to be "highly qualified."

Debray (2005) argued that Bush took Clinton's education proposal and was able to get them passed. He did this by taking advantage of a shift in the dynamics of congress following the 2000 elections that made members more willing to work together. He also took advantage of the united congress that followed the attacks of September 11, 2001. Debray (2005) also argued that Bush was willing to move toward the center, broadening the range of proposals that he would accept, and to drop the insistence that vouchers be included in the new law. Finally, she claimed

that the final vote represented Republican loyalty to a Republican president during a time of national crisis.

According to McGuinn (2005), due to the converging political opinions over standards and accountability as mechanisms for reform, NCLB (2002) won broad bi-partisan support in both the house and senate. This law continued the shift from the original ESEA which assumed that most schools were doing okay but that schools needed to better serve certain segments of the population, to the assumption that all American schools needed to improve the quality of education provided to students.

Although McGuinn (2005) argues that there was an evolution of political thought among both Republicans and Democrats that led to the convergence of their respective reform agendas and ultimately resulted in the passage of NCLB (2002), others, such as Kohn (2004b) have argued that NCLB (2002) was part of a conservative agenda of devolution and market based reforms, intended to ensure that all schools would ultimately fail. According to this reasoning, such massive public school failure would make the more traditional Republican position on education reform more palatable to the American public.

NCLB (2002) codified testing for accountability. Under NCLB (2002), children were to undergo annual testing in language arts and mathematics in grades three through five, eight, and once in high school. The intent of these assessments was to provide schools with information that could be used to improve instruction, and to hold schools and educators accountable for educating all students. Each school was expected to educate each child to proficiency or above in language arts and mathematics, by the year 2014. States set intermediate "AYP" targets for the number of students expected to score above the cut score for proficiency on the state

assessments. Schools that did not have a sufficient percentage of students scoring above these cut scores for two consecutive years were labeled “in need of improvement.”

NCLB (2002) prescribed a series of progressively more severe sanctions for schools designated “in need of improvement.” After two years of failing to make AYP, schools were required to allow students to transfer to another school in the same district that made AYP, even if the receiving school was a charter school. The district was required to pay transportation costs. Consequently, schools “in need of improvement” lost state per pupil funding at the same time that the district was required to pay these additional transportation costs.

After three consecutive years of failing to make AYP, schools were required to provide Supplemental Educational Services (SES) to low income students attending the school, in addition to continuing to offer parents the option of transferring their children to a higher performing school within the district’s boundaries. The costs of SES placed additional financial burdens on schools designated “in need of improvement.”

After four consecutive years of failing to meet AYP targets schools were labeled in need of “corrective action.” Schools in need of “corrective action” were required to do one of the following: replace all staff members; adopt a new curriculum; or lengthen the school day or year. Schools in “corrective action” were required to continue to offer SES and the option to transfer to another school in the district.

Finally, schools that failed to make AYP five years in a row were required to initiate restructuring. Restructuring could include closing the school, turning it into a charter school, or turning it over to a private management company.

In addition to the consequences imposed on schools and educators for poor student test scores, NCLB (2002) included provisions that required districts to provide additional supports to

schools that struggled to meet APY targets, and that rewarded schools that demonstrated progress.

The testing for accountability debates began with the wave of criticism initiated by the publication of *A Nation at Risk: The Imperative for Educational Reform* (Commission for Excellence in Education, 1983). Proponents of testing for accountability included public education critics who were members of George H. W. Bush's education cabinet. These included David Kearns (Kearns, 1988; Kearns & Doyle, 1989), Chester Finn (Finn, 2002; Finn, Kanstoroom, Rothstein, & Honig, 2001; Finn, Manno, & Vanourek, 2001; Finn & Ravitch, 1996; Kanstoroom and Finn, 1999), and Dianne Ravitch (1993, 1994, 1996; 1999; Finn & Ravitch, 1996). Other critics of public education included Eric Hanushek (1989, 1994, 2003 2005; Hanushek & Raymond, 2005) and Herbert Walberg (1994, 1998). Finally, these proponents also included advocates of systemic school reform such as Marshall Smith and Jennifer O'Day (1991). In addition, the perspective of advocates for minority groups who supported standards, testing and accountability reforms will also be presented (Archer, 2006; Reid, 2005; Rury, 2012; Salzman, 2006)

Opponents of testing for accountability included public school defenders such as David Berliner and Bruce Biddle (1995) and Gerald Bracey (1987, 1996); neo-progressive educators such as Susan O'Hanian (1999, 2000, 2003), and Alfie Kohn (2000; 2004a, 2004b); psychometricians including Daniel Koretz (2008) and George Madaus (Madaus, 1988; Madaus, Russell & Higgins, 2009); researcher at the economic policy institute, Richard Rothstein (Finn, Kanstoroom, Rothstein, & Honig, 2001, Rothstein, 2004, 2008; Rothstein, Jacobson & Wilder, 2008); professor of educational research, Peter Airasian (1987); and educational theorist Michael Apple (2001).

Conclusion

Intelligence tests and accountability tests both emerged from a desire to impose objective order on a disorderly world (Wiebe, 1967). During the first 30 years of the 20th Century, this meant employing objective science to make overcrowded schools run more efficiently, providing opportunities to historically marginalized people, reconstructing society by placing people in the niche they were best suited to fill, and rationalizing the existence of inequality. Similarly, testing for accountability emerged during the last 40 years of the 20th century as a means to improve education in general, and improve equity for marginalized groups in particular. Accountability testing as a reform was based upon standardizing and increasing the rigor of the curriculum, using objective measures to determine whether the students had learned the content, and utilizing rewards and punishments to motivate educators to improve instruction.

This chapter explored the historical events, popular political orientations, scientific discoveries and educational trends that led to the ascendancy of intelligence testing during the progressive era, the resurgence of the intelligence testing debates in the late 1960s and mid-1990s, and proposals to use achievement tests to hold educators accountable during the 1980s, 1990s, and early 2000s, ultimately leading to the passage of NCLB (2002). The next three chapters will include a description and analysis of the arguments from both sides of the respective testing debates. This will begin with Chapter 3, and the analysis the intelligence testing debates of the 1920s.

Chapter 3: The Intelligence Testing Debates

Introduction

The debates surrounding intelligence testing were the product of the publication of the data and interpretations from the World War I Army Intelligence testing program (Brigham, 1923; Yerkes, 1921), as well as published works that pertained to the use of intelligence testing in schools, but which preceded the publication of the results of the Army Tests (Goddard, 1922; Haggerty, et al., 1920; Terman, 1919, 1920; Thorndike, 1919). These early promoters of intelligence testing included Lewis Terman (1919, 1920, 1922a, 1922b, 1923; Haggerty, et al., 1920), Henry Herbert Goddard (1922), Edward L. Thorndike (1919; Haggerty, et al., 1920) and Guy M. Whipple (1922; Haggerty, et al., 1920).

In 1922 and early 1923, Walter Lippmann and Lewis Terman debated each other in a series of articles published in The New Republic (Lippmann, 1922a, 1922b, 1922c, 1922d, 1922e, 1922f, 1923; Terman, 1922b). These debates were oriented around appropriate interpretations of the results from the Army testing program, the inherent nature of intelligence, the validity of the tests as measures of inherent intelligence, and the appropriate use of the tests. John Dewey (1922/ 2008a, 1922/ 2008b) briefly joined these debates with two articles published in The New Republic, in which he raised concerns regarding intelligence tests and their implications for both education and democracy.

William Bagley (1922a, 1922b) also exchanged arguments with Terman (1922a) and Whipple (1922) in 1922. Like Lippmann, Bagley also questioned the science on which the tests were based (Bagley, 1922a, 1922b, 1925; Minton, 1987/ 1990). In addition, the Bagley and Terman debate centered on the assumptions made by psychologists about the limitations imposed by innate intelligence, and the implications of these limitations for education in a democracy.

Finally, both Bagley (1925) and Horace Mann Bond (1924a, 1924b) responded to Carl Brigham's *A Study of American Intelligence* (1923), wherein Brigham interpreted the results from the Army tests as demonstrating a racial hierarchy based on inherent merit. Both Bond and Bagley focused on the very different environments that immigrants and minorities experienced compared to their native white peers (in this context, "native" refers to people born in the United States).

This chapter provides an analysis of the intelligence testing debates. These debates were addressed by topic. These topics included the nature of intelligence, the relationship between inborn intelligence and "race," the validity and precision of the tests, the use of the tests to treat humans and human institutions like objects, the proper role of intelligence testing in education, and the nature of democracy. The analysis of the arguments from these debates will be used to begin to identify emerging themes regarding testing policy in public education.

The Nature of Intelligence

Tests to measure intelligence were developed before psychologists had agreed on a common definition of just what was being measured (Bagley, 1922; Bond, 1924a, 1934/ 1966; Lippmann, 1922a, b, f). Lippmann (1922a, 1922b) argued that the tests were based upon a number of abstract processes that a small number of test developers guessed were components of intelligence. He also questioned whether it was possible to measure an abstract construct that had not yet been properly defined (Lippmann, 1922f).

While there was no common definition of intelligence, several of the proponents of IQ testing proposed possible definitions. These included the ability to learn (Bagley, 1922a; Haggerty, et al., 1920, p. 27); that which separated children who succeeded in school from those

who were less successful (Terman, 1922b); and the ability to adapt to the social environment (Goddard, 1922). Critics of intelligence testing also proposed possible definitions. For example, Lippmann suggested that intelligence was "the capacity to deal successfully with the problems that confront human beings," (Lippmann, 1922b, p. 246).

However, this lack of a common definition of what exactly they claimed to be measuring did not trouble the intelligence testers (Goddard, 1922; Whipple, 1922). Indeed, some argued that it was not necessary to define a thing in order to measure it. Electricity and light had both been measured before scientists could define these constructs. The new tests of intelligence might even provide insight that would help psychologists to flesh out a proper definition of intelligence (Goddard, 1922; Whipple, 1922). To be sure, E. G. Boring (New Republic, June 6 1923) offered a somewhat circular definition of intelligence as being whatever it was that intelligence tests measured.

Despite this lack of common definition of intelligence, testers believed that it was a quality inherent to the individual that determined his or her merit or talent. Indeed, most of the participants in the debates, including critics of intelligence testing, accepted that at least a portion of intelligence could be attributed to genetic endowment (Bagley, 1922a, 1922b; Brigham, 1923; Goddard, 1922; Lippmann, 1922f; Terman, 1919, 1922a; Whipple, 1922). What differentiated the proponents of IQ testing from their critics was the degree to which they believed that genetic endowment was the origin of intelligence. Most of the critics of intelligence tests believed that the environment and experience played a critical mediating role in shaping individual intelligence (Bagley, 1922a; Bond, 1924a, 1924b; Lippmann, 1922e, 1922f). In contrast, promoters of intelligence testing tended to believe that intelligence was predominantly inherent,

and therefore, little could be done to alter it (Brigham, 1923; Goddard, 1922; Terman, 1919, 1922; Whipple, 1922).

Promoters of intelligence testing provided a variety of evidence supporting their belief that intelligence was inherent. Much of this evidence was based upon correlations, assumptions and personal bias. For example, Francis Galton (1870) found that prominent citizens were disproportionately related to other prominent citizens. Similarly, Terman (1919) found that children with higher IQ scores were more likely to be related to prominent citizens. Terman (1919) also pointed to the high correlations between IQ scores of parents and children, as well as between siblings.

Studies of paternal and fraternal twins were also used to provide evidence that natural endowment was more important than the environment in determining intellectual capacity. Terman (1922b) pointed to the higher correlation of IQ scores between monozygotic twins than between dizygotic twins. Because monozygotic twins shared identical genetic material, it was assumed that this higher correlation between the IQ scores of monozygotic twins than between dizygotic twins indicated that intelligence was primarily inherited.

Occasionally, evidence that seemed to support a strong environmental influence on intelligence was used to support *a priori* beliefs regarding the genetic origins of intelligence. For example, Terman (1916) "tested twenty children in an orphanage and found only three that were fully normal," (Lippmann, 1922e, p. 330). Arguing that the orphanage in question was relatively good, and therefore provided an environment that was equivalent to the home life of a typical middle class family, Terman concluded that the majority of the orphans were genetically inferior.

Similarly, Brigham (1923) explained the correlation between intelligence test scores and years of schooling by assuming that more intelligent individuals were more successful in school,

and therefore remained in school longer than less intelligent individuals. Both Goddard (1922) and Terman (1919) also attributed school persistence to success caused by inherent intelligence rather than attributing high intelligence test scores to school persistence. By contrast, Bond (1924a, 1924b) argued that the discrepancy between a typical school term for schools serving different populations contributed to the variations in average intelligence between these populations.

Likewise, Terman (n.d.) and Lippmann (1922d) attributed the decreasing correlation between SES and intelligence as children grew older to opposite causes. Terman argued that this was proof that intelligence was genetically determined. If intelligence was caused by the environment, then one would expect that the correlation between IQ and SES would increase the longer the child was exposed to parental influences.

Conversely, Lippmann (1922d) concluded that as a child grew older, his or her circle of influence expanded. Therefore, one would expect the correlation between social status and IQ to decrease as the child was exposed to a greater variety of environmental influences.

In addition, Lippmann (1922e) wrote that if, as psychologists claimed, the growth curve of intellectual development was very steep in infancy, the time when the environment would have its greatest impact on intelligence was during this developmental period. It should not, therefore, be surprising that there was such a high correlation between parental social status and the IQ of the very young child, since the child's primary environmental exposure was the home.

However, promoters of the new intelligence tests countered this argument by suggesting that intelligence followed a growth trajectory similar to that of other physiological structures. Genetically pre-programmed growth in the physiological structures associated with the nervous system caused the steady growth in mental age from infancy through middle adolescence

(Goddard, 1922; Terman, 1919, 1922a; Whipple, 1922). Just as individuals stopped growing taller as they reached middle adolescence, they also stopped growing in their intellectual capacity. Indeed, Terman (1919) argued that the public should be just as willing to accept that genetic endowment caused people to vary in intelligence and that intelligence stopped growing at a genetically predetermined point, as they were to accept that genetic endowment caused people to vary in stature and that people stopped growing taller at a genetically predetermined point.

Indeed, both Goddard (1922) and Terman (1919) expressed the belief that individual intelligence could be decreased through disease, accident, or injury. Certainly, there was some evidence to back up this claim. Terman (1919) referred to a number of studies which demonstrated that IQ changed little between the ages of four and fifteen years of age.

Terman (1922a) supported this assertion that intellectual capacity stopped growing at a predetermined point with data from the Army Intelligence Testing Program. This data showed that the average intelligence for men at different ages did not increase between the ages of 21 and 31. However, Bagley (1922a, 1922b) described the idea that intellectual capacity suddenly stopped growing as purely hypothetical. He noted that Terman's claim that IQ did not grow between the ages of 21 and 31 was based on a cross-sectional sampling of different men at varying ages, and not a longitudinal study following the same cohort of men as they grew older (Bagley, 1922b).

Like Bagley, Lippmann (1922f) also challenged the claims of stability of IQ, noting that IQ was a standardized score, and therefore represented an individual's position relative to a norm referenced group of the same age. To argue that IQ did not change gave the false impression that there was no change in absolute intelligence. What actually showed little change was one's

position relative to others in his or her age cohort. Lippmann concluded that IQ scores masked growth in intellectual capacity at different stages of the child's development.

While much of the evidence seemed to support the notion that intelligence followed a growth trajectory and then stabilized, there were some studies that showed that intelligence scores tended to increase if five or more years had passed between the original test and the post test. However, Terman (1919) dismissed this evidence, suggesting it was caused by different forms of intelligence tests. While it is certainly possible that different versions of intelligence tests would have caused the difference in IQ scores observed over time, the possibility that intellectual capacity relative to others might actually improve in some people was not even entertained.

Terman (1919) also used the correlations between intelligence test scores and SES or parental occupation to support his belief that intelligence was determined by innate endowment. He noted that children with high intelligence test scores often had parents who made more money or who held a professional occupation. However, as Bond (1924a) pointed out, this was based on the assumption that intelligence determined SES or choice of profession, since the parents' IQ had not been tested. Bond (1924a, 1924b) argued that this relationship could also go the other way. Indeed, he provided evidence that showed a relationship between living conditions as well as per capita spending on schools to intelligence test scores.

Bagley (1925, chapter IV) also found strong correlations between the average score on the army alpha among white males from each of 26 states and the quality of schools in those same states. Again, this supported the important role of experience, particularly of schooling, in influencing scores on the Army Alpha intelligence test.

Race

Related to the belief that genetic endowment was the primary cause of intelligence was a belief that differences in average intelligence test scores for members of certain "racial" and ethnic groups were due to the genetic factors that determined "race." First generation psychologists tended to come from a middle class, white protestant background (Cravens, 1978/ 1988). They were immersed in a cultural context in which middle class white Protestants felt increasingly threatened by immigrants from Southern and Eastern Europe, as well as African Americans. Furthermore, they were heavily influenced by the emerging science of evolution, and many believed that different people of different "races" and ethnicities amounted to a living fossil record of human evolution (Reed, 1987/ 1990). As such, they believed that "races" and ethnicities could be ranked by average intelligence test scores into an evolutionary hierarchy (Cravens, 1978/ 1988). For example, in arguing that the Army intelligence tests were accurate measures of intelligence, Goddard offered as proof the fact that many of the enlisted men with low IQs were "foreign," (Goddard, 1922, p. 26). In the context of this statement, Goddard seemed to have assumed that it was common knowledge that "foreign" was synonymous with "inferior."

Brigham (1923) concluded that the evidence from the World War I intelligence testing program showed conclusively that there was an evolutionary hierarchy of "races." Using intelligence test scores as the indicator, Brigham ranked the "races" from superior to inferior as follows: Nordic, Alpine, Mediterranean, Negro. This ranking also happened to coincide with a popular contemporary belief in Nordic supremacy (See for example Grant, 1916).

Some of the evidence that confronted Brigham appeared to contradict these findings. For example, Northern Blacks had higher intelligence test scores than did Southern Blacks even after

controlling for years of schooling. Brigham (1923) explained this by arguing that Northern Blacks were genetically superior to Southern Blacks. This was because Northern Blacks were more likely to have "a greater amount of admixture of white blood," (Brigham, 1923, p. 192). As such, these more intelligent Blacks were capable of understanding that living conditions were better in the North, and were therefore more likely to migrate.

Similarly, results from the Army testing program showed that immigrants who had been in the United States longer had higher intelligence test scores than did more recent immigrants. Rather than admit that the tests might measure something other than inherent intelligence, Brigham (1923) argued that the genetic material of the more recent immigrants from Southern and Eastern Europe was inferior to that of earlier immigrants from so-called Nordic countries.

However, Bond (1924b) demonstrated that when Blacks were provided with better educational opportunities than whites, Blacks outperformed Whites on intelligence tests. Using data from the Army Intelligence Testing Program, he demonstrated that the mean intelligence test scores of both Black and white enlisted men correlated highly with Ayers's state rankings of schools (Ayers, 1918). Bond showed that the African American recruits from the two highest scoring states outscored the whites in the four lowest scoring states.

Later, Bond (1934/ 1966) would argue that Brigham's use of "years in school" as a measure of education, especially when comparing white and Black soldiers, was inadequate. He noted that a year of schooling was very different for Blacks than it was for whites. Blacks attended school on average only five months during the school year, whereas whites attended on average seven months. There were also significant differences between the quality of the facilities and the training of the teachers of the schools attended by Black children as opposed to the schools attended by white children. Black soldiers had frequently been raised in rural areas,

isolated from educational and cultural experiences. Furthermore, many of the Black soldiers had grown up in conditions of poverty. Finally, many of the Black soldiers were illiterate, rendering them less likely to have had access to the collective knowledge of the white American culture. As such, “years of schooling” was an inadequate measure of the sum of an individual’s educational experiences. Bond concluded, "…. that is the indisputable truth that Alpha measures environment, and not native and inherent capacity. Instead of furnishing material for the racial propagandists and agitators, it should show the sad deficiency of opportunity which is the log of every child, white or black, whose misfortune it is to be born and reared in a community backward and reactionary in cultural and educational avenues of expression,” (1924b, p. 201).

Likewise, Bagley (1925) challenged Brigham’s conclusion that the lower mental age scores of more recent immigrants was due to changing immigration patterns by pointing out that those immigrants who had been in America longest had most likely attended American schools. Furthermore, he suggested that the more recent immigration came from countries that had less well developed education systems. Indeed, he argued that there was a fairly high correlation ($r=0.84$) between the Alpha and Beta scores of recent immigrants and the ranking of the schools of their nation of origin.

Finally, Bond (1924a, 1934/ 1966) repeated Thomas Garth’s (1921 cited in Bond, 1924a and in Brigham, 1923) axiom that racial comparisons were not valid unless the representatives being tested shared the same developmental experiences. As Bond (1924a) pointed out, the intelligence testers assumed, "… that the minimum of experience possessed by a Negro from the horribly inefficient schools of the far South places him on a plane of equality, for purposes of comparison, with the graduate of the highly standardized grammar school systems of California or the District of Columbia. They assume that the experience gained by a Negro living in the

slums of Memphis is sufficient to warrant comparison with the product of the proudest scions of Malden or of Beverly Hills," (p. 198).

Brigham (1923) circumvented this argument by contending that to draw a sample from populations that shared the same environment would guarantee a biased sample. His reasoning was based on the assumptions that SES was determined by native intelligence, and that the average Black child could not keep up with the average white child. Therefore, to draw a sample of children matched for their level of educational achievement was to draw a sample that either disproportionately sampled the more intelligent Black children, or disproportionately sampled inferior white children. Even assuming a causal relationship between inherent intelligence and economic status, this ignored the fact that the average Black child was denied access to the same quality of education as was enjoyed by the average white child.

Tests as Precise Measures of Innate Intelligence

Barely 12 years after Goddard introduced the Binet-Simon test to the United States, a number of psychologists asserted great confidence in the precision and validity of these new tests (Goddard, 1922; Terman, 1919; Whipple, 1922). To Goddard, "Testing intelligence is no longer an experiment or of doubted value. It is fast becoming an exact science," (1922, p. v ii). Whipple concurred, stating, "… our intelligence tests do measure, with a precision that is surprisingly satisfactory, a factor which is of the utmost significance for educational progress," (1922, p. 601).

The confidence which testing advocates spoke of the intelligence tests was not limited to individual intelligence tests. They also praised the accuracy of the new group intelligence tests first developed for the Army Intelligence testing program. Indeed, despite acknowledging a

number of methodological problems with the Army Intelligence Tests, which included the inconsistent adherence to the testing protocol and the lack of consistent criteria for determining which men qualified for the Beta test (Brigham, 1923; Gould, 1981/ 1996;), both Goddard (1922) and Brigham (1923) declared the group tests to be valid measures of intelligence.

Promoters of intelligence tests offered a variety of evidence as proof of the validity of the tests. Often, they relied on the high correlation between intelligence test scores and other measures of "success." Such measures included SES, occupation, and parental occupation (Brigham, 1923; Terman, 1919; Terman, 1916 cited in Minton, 1987/ 1990; Thorndike, 1919). Other correlation evidence was also offered as proof of the validity of the tests. Some of this evidence was based on assumptions about different population groups, or the relationship between intelligence and school persistence. For example, IQ scores of foreign born school children correlated with the subjective judgments of teachers regarding the intelligence of these children (Terman, 1919). This assumed that teachers' judgments were based on factors related to intelligence, and not on the language proficiency of individual students or personal bias against immigrants on the part of the teachers. Lippmann (1922c) also noted the correlation between teacher judgment and intelligence test scores. However, he suggested that there was no evidence that teacher judgment was any more valid than an intelligence test as a measure of intelligence.

Goddard (1922) suggested that the large number of illiterate enlisted men who performed poorly on the Army tests was evidence of the validity of the tests. This was based on Goddard's unquestioned assumption that illiteracy was caused by low intelligence.

Furthermore, Brigham (1923) used the high correlation between IQ scores and "years of schooling" as evidence that the tests measured intelligence. This was based on the assumption that intelligent people persisted in school longer than less intelligent people, and that students

who dropped out of school did so only because they lacked the intelligence necessary for school success.

Intelligence tests also tended to correlate with other tests of intelligence (Brigham, 1923; Lippmann, 1922c; Yoakum & Yerkes, 1920). Although he acknowledged the high correlation between different tests of intelligence, Lippmann noted that this only indicated that the tests measured "the same capacities," (1922c, p. 276). He questioned, however, whether the tests measured an individual's "ability to deal with life," or merely his or her capacity to pass a test or to perform well in school. He pointed out that the tests were still so new that psychologists had not yet been able to do follow up studies to determine whether intelligence tests predicted success in life.

In addition, Lippmann (1922c) noted that some individuals were highly motivated to perform well in the testing environment, while others performed poorly in the testing environment. Under these circumstances, the test was more a measure of one's motivation than of inherent intelligence.

Elsewhere, Lippmann (1922b, 1923) suggested that process for determining intelligence scores was somewhat arbitrary. Mental ages were determined by the proportion of test items that an individual of a given age could successfully complete, and this was compared to the number of items that an arbitrary number of individuals of the same age could successfully complete. In addition, arbitrary decisions affecting testing procedures also determined how individuals were classified. For example, the addition of time limits to the procedures of the Army Alpha Tests had the effect of limiting the number of enlisted men who tested as "A men," (Lippmann, 1922b). Terman (1922b) was quick to point out that time limits did not affect the rank order of the men. However, as Lippmann (1923) responded, procedures such as time limits determined

the percentage of men who fell into the different categories of intelligence, and determined whether individual men were classified as "A" men or "B" men. Such decisions could have high stakes for individual men during a time of war.

Finally, Lippmann (1922c) argued that just because the intelligence tests were a good measure of school success, and school success was a good indicator of life success, this did not mean that the tests were a good measure of the capacity to deal successfully with life. This was because the tests might simply be measuring school learning, and not inherent ability. Lippmann concluded that while the tests might be valid for ranking individuals or placing students in grades, there was not enough evidence to conclude that they were valid measures of absolute inherent intelligence (Lippmann, 1922d, 1922f).

Intelligence Tests and Testers Objectifying Humans

IQ tests were, or appeared to be, objective measures. This objectivity combined with the desire of first generation psychologists to establish psychology as a legitimate science (Cravens, 1978/ 1988; Gould, 1981/ 1996; Murdoch, 2007) contributed to the tendency of the IQ testers and their tests to *objectify* people. In this context, *objectification* referred to the treatment of human beings as though they were things (Friere, 1990). IQ testers treated people as though they were objects by using IQ and mental age scores as though synonymous to an individual's intelligence or value as a human being, by using language that referred to humans as objects, or by drawing analogies between human beings and objects.

IQ and mental age scores were an important component in the attempt to establish psychology as a legitimate science on par with more established sciences (Gould, 1981/ 1996; Murdoch, 2007), a "science of man" (Cravens, 1978/ 1988). Humans were animals, and had

evolved like other animals. They were therefore subject to natural law (Cravens, 1978/ 1988). Many psychologists believed that animals and humans could be rank ordered along an evolutionary scale, and that intelligence tests offered a way to objectively measure the quality of individuals and groups of people (Cravens, 1978/ 1988; Goddard, 1922). This quality was synonymous with merit.

Lippmann (1922a, 1922c) noted that a mental age or IQ score was a numeric summarization of an individual's performance on a collection of tasks believed to measure intelligence (Lippmann, 1922a, 1922c). Using a number to represent the abstraction of "intelligence" was an example of reification, the logical fallacy of treating an abstract construct as though it were a thing (Gould, 1981/ 1996).

Additionally, some critics of intelligence testing argued that the use of a numeric score gave the false impression that intelligence existed in an absolute quantity, and that the quantity of intelligence that any given individual possessed could be precisely measured (Gould, 1981/ 1996; Lippmann, 1922b, 1922f). Both mental age and IQ scores gave the false impression that these measures were on continuous scales with each point representing a unit of intelligence. Instead, mental age was merely a classification based upon the average age at which the reference group could correctly complete a given number of tasks. IQ scores were standardized scores, representing an individual's position relative to others of the same age. Neither mental age nor IQ scores represented an amount of something (Lippmann, 1922b, 1922f). Referring to the intelligence test, Lippmann wrote, "It does not weigh or measure intelligence by any objective standard. It simply arranges a group of people in a series from best to worst by balancing their capacity to do certain arbitrarily selected puzzles, against the capacity of all

others. The intelligence test, in other words, is fundamentally an instrument for classifying a group of people," (Lippmann, 1922b, p. 247).

IQ and mental age scores were used to justify the treatment of individuals with disabilities as though they were defective or even sub human (Goddard, 1922; Jackson & Weidman, 2004/ 2006;). The correlation between intelligence test scores and SES enabled psychologists to attribute wealth to the greater inherent intelligence of the wealthy, even as they blamed poverty on the lower inherent intelligence of the poor (Brigham, 1923; Goddard, 1922; Terman, 1916, 1919). Racial and ethnic groups were ranked by average intelligence test scores, providing "evidence" that some groups were inferior or sub-human (Brigham, 1923; Goddard, 1922; Terman, cited in Bond, 1924b; Terman, 1916; Terman, 1919; 1922b). Such rankings were used to rationalize inequitable and often inhumane treatment of members of these groups. In addition, intelligence test scores were used to justify restrictive immigration policies and eugenics laws allowing for the sterilization, and in the case of Nazi Germany, the extermination of "defective" humans and members of "inferior" racial and ethnic groups (Cravens, 1978/ 1988; Jackson & Weidman, 2004/ 2006; Murdoch, 2007). Intelligence test scores placed artificial limitations on children (Bagley, 1922a; 1922b, 1925; Lippmann, 1922d). Finally, they used numbers to designate some humans as having less value than others (Bond, 1924a, 1924b). By assigning a numeric value to humans, intelligence tests in effect treated them as though they were objects.

In addition to assigning worth to individual human beings, intelligence test scores were also used to place a value on groups of people in a "racially" and economically ordered social caste system. Bond (1924a, 1934/ 1966) suggested that IQ tests were just another in a long line

of strategies used to rationalize the inhumane treatment of one group of people by the dominant group. In this way, intelligence tests were a tool of hegemony.

Not only did they encourage people to be treated as though they were their scores on intelligence tests, first generation psychologists frequently used language in ways that referred to human beings as though they were objects. These early psychologists often borrowed language or drew analogies from the more established disciplines such as engineering in order to evoke the impression that psychology was scientifically rigorous and used precise measurement instruments (Goddard, 1922; Spearman, 1904; Terman, 1919, 1920). Referring to humans as though they were objects also gave the impression that psychology was a dispassionate, objective discipline.

Indeed, there was a relationship between "objectivity" in science, and the objectification of humans. Objectivity suggested that one's judgment was not swayed by emotion, but only the facts. Therefore, in science, as in war, there were advantages to objectifying subjects. Several of the psychologists referred to humans in language borrowed from other sciences. Charles Spearman (1904) consistently used the word "reagent" to refer to the human subjects participating in his studies.

Others drew analogies from other disciplines, such as medicine or engineering, when discussing measurement of human intelligence. Terman (1919) wrote, "Standardization is coming to play the same role in psychology that it has long played in the various branches of applied science. The architect or bridge engineer plans his structure with constant reference to foot-points of strain which various materials will withstand. The physician analyzes a drop of blood and, by comparison of corpuscle count and haemoglobin [sic] with the norms for health and disease, is able to render an important diagnosis. The psychologist working with mental tests

may be compared with the palaeontologist [sic] who finds in a gravel bed of some prehistoric age a skull cap, a fragment of jaw, and a broken humerus," (p. 4). Again, this appeared to be an attempt to present psychology and intelligence testing as legitimate sciences. However, there were other examples from the samples of Terman's writings where he used language that referred to humans as objects.

Another form of objectification is the commodification of people. Intelligence testers frequently used language that described humans as economic or natural resources. Terman (1919, 1920) described people in terms of the "quantity" of intelligence that they possessed. He referred to school children who were in the first grade as "the raw material with which the school is to work," (1919, p. 42; 1920, p. 22). He compared the new intelligence tests to an assay of "how much gold is contained in a given vein of quartz," (Terman, 1919, p. 1).

Indeed, early psychologists frequently referred to humans as "material" (Thorndike, 1919, p. 56) "human material" (Goddard, 1922, p. 29) or "raw material" (Terman, 1919, p. 42; 1920, p. 22). They wrote of using this material efficiently (Goddard, 1922; Thorndike, 1919) and conserving the nation's intellectual assets (Terman, 1919, p. 288). They wrote of children as being "inferiors," "burdens," "assets," (Terman, 1919, p. 132-133), "defectives," (Goddard, 1922 pp. 18, 77; Haggerty, et al., 1920, p. 27; Terman, 1919, p. 285;) and "superiors" (Terman, 1919, p. 190).

Both Goddard (1922) and Thorndike (1919) suggested that humans were resources that should be efficiently distributed and utilized. This was what Goddard (1922) meant when he drew an analogy between intelligence testing for social engineering and the engineer choosing materials for a bridge. He wrote, "The mechanical engineer could never build bridges or houses if he did not know accurately the strength of materials, how much of a load each will support. Of

how infinitely greater importance is it that we should know the strength of our materials," (Goddard, 1922, p. 29). Later, he again evoked comparisons to engineering when he stated, "It is a maxim in engineering that a bridge is not stronger than its weakest part. The same is largely true of society. It must be understood however, that weakness is not determined by the size of the part but by the relation the size or strength of the part bears to the work it has to do," (Goddard, 1922, pp. 34-35). Here, he used dehumanizing language to justify the social engineering proposals that were incorporated throughout this series of lectures.

Similarly, Goddard wrote, "Intelligence is the potentiality of the machine. Knowledge is the material upon which it works. Knowledge is the raw material. Intelligence determines what we do with it," (Goddard, 1922, p. 8). Finally, referring to the potential for using the new tests to assist in social engineering, Goddard wrote, "Whether we are thinking of children or adults it [the intelligence test] enables us to know a very fundamental fact about the human material," (Goddard, 1922, p. 29).

Thorndike (1919) also suggested that humans were resources to be efficiently distributed when he suggested that placing intelligent people in occupations that required little intelligence created a “large unused surplus of intellect," (p. 55). Referring to the need for the army to use the intelligence tests to efficiently fill occupations, he wrote that “each [occupation] had to be filled so as to leave the best possible material to fill every other requisition," (p. 56).

John Dewey (1922/ 2008b) attributed this habit of referring to human beings using dehumanizing language to industrialization and the habit of science of using the language of statistical averages. “Our mechanical, industrialized civilization is concerned with averages, with percents,” (p. 295).

Intelligence Testing In Education

If the Army tests seemed to prove anything, it was that large numbers of people could be tested, classified, and sorted efficiently (Thorndike, 1919). Psychologists saw an opportunity to promote the new group intelligence tests in schools (Terman, 1920). As Terman stated, "It is becoming clear, however, that their [intelligence tests] greatest usefulness will be found in their universal application to school children…'A mental test for every child' is no longer an unreasonable slogan," (Terman, 1920, p. 20).

Psychologists believed that the tests could not only make schools operate more efficiently, the use of tests in schools could also serve a broader social efficiency function. Indeed, the authors of The National Intelligence Test (Haggerty, et al., 1920) listed the four uses for intelligence testing in schools for which promoters of intelligence testing typically advocated. Each of these was oriented toward making either the school or society more efficient. They included 1) sorting students into instructional groups based on ability; 2) identification of children with disabilities; 3) identification of gifted children for special instruction; 4) and, guiding students toward vocations matched to their abilities.

Testing advocates, and many critics of intelligence testing, believed that the tests should be used to place students into homogeneous ability groups (Goddard, 1922; Lippmann, 1922d, 1922f; Haggerty, et al., 1920; Terman, 1919, 1920). Students would be placed in classes according to their mental ages rather than their chronological ages. This would individualize education for the gifted and unintelligent alike, avoid frustrations of moving too fast or slow, and provide an efficient use of resources (Goddard, 1922; Terman, 1919). Furthermore, enormous amounts of time and resources could be saved if curriculum and instruction were matched to the ability of the students (Whipple, 1922).

Several of the testing advocates argued that the content taught in schools was too difficult for some students (Goddard, 1922; Terman, 1919, 1920). Students should be taught content that they were capable of mastering. Therefore, Terman (1920) proposed parallel educational tracks for the intelligent and unintelligent students, respectively. The content of these tracks should differ according to the ability of the students.

Indeed, Terman (1919) called for the end of the practice of allowing university entrance exams to drive educational standards for high schools. Rather, there should be alternative courses of study that would provide a useful education to those students who lacked the ability to successfully attend college.

Testing advocates also suggested that compulsory attendance laws might be inappropriate for the less intelligent (Goddard, 1922; Terman, 1919). They questioned whether children should be forced to continue to attend school once the content of the curriculum was beyond their capacity to benefit. At times, both Terman (1919) and Goddard (1922) went so far as to argue that students should not be allowed to continue in education once the content was beyond the limits of their ability. Both argued that students should be required to achieve minimum IQ scores to be allowed to continue with high school or college work.

In addition to individualizing instruction, tests could also be used to identify children who, in the language of the times, were feebleminded or defective (Goddard, 1922; Terman, 1919). However, this was not intended to provide these students with targeted instruction that would catch them up with their "normal" peers. Rather, it would allow the feebleminded to be funneled into treatment that would prevent them from becoming promiscuous, criminals or delinquents. It also allowed them to be trained for useful work that was within their capabilities (Goddard, 1922; Terman, 1919).

Identification of the feebleminded would also allow them to be placed into institutions or colonies, thereby isolating them from the normal population. This would make classrooms more efficient, allowing teachers to dedicate time and effort to the children that could actually benefit from school. Furthermore, the mentally defective children would not slow the progress of the "normal" and "superiors." In addition, it would prevent scarce resources from being wasted on attempts to teach the feebleminded a curriculum for which they lacked the capacity to benefit, resources that could be better spent on "normal" and "superior" children (Goddard, 1922; Terman, 1919; Whipple, 1922). Finally, it would spare "normal" and "superior" children from having to sit next to "defective" children (Goddard, 1922).

Bagley (1922a) countered that schools could not simply write off children because they lacked a certain level of innate intelligence. He expressed his belief that children whose intelligence was low, but above a minimum threshold, could benefit from education. Furthermore, any such gains, no matter how slight, benefited the nation. The cynical suggestion that teachers could do little for children with low IQs disempowered teachers. He wrote, "As I watch these teachers at their work it is not what they can not [sic] do that impresses me, it is rather the miracles that their consummate art enables them to perform. I have seen dull eyes lighted with a momentary gleam of intelligence. It was a little light in a world of darkness…A little more light for the common man this year, next year, a hundred years from now, and the battle for humanity, for democracy, and for brotherhood is won," (Bagley, 1922a, p. 384).

Terman (1922a) sarcastically responded to Bagley, arguing that Bagley denied scientific proof that there were innate limits to each child's ability. He suggested that Bagley would prefer to believe in "miracles" than to accept scientific evidence, accusing him of using "Christian Science Psychology," (Terman, 1922a, p.58). Terman alternately accused Bagley of denying

science, embracing mysticism, and allowing sentiment to obstruct his objectivity. In one such passage, Terman wrote, “But almost immediately his vision is blurred by the moist tears of sentiment and the eloquent address closes with a rhapsodic peroration on the miracles that skillful teachers work with morons and on the ultimate illumination of the world by gleams of light struck from dull minds,” (p. 58). This passage was representative of Terman’s responses to his critics, in which he frequently resorted to sarcasm and ridicule to counter arguments against his beliefs about intelligence and testing.

Lippmann (1922c, 1922d) and Dewey (1922/ 2008a) both argued that it might be useful to use the tests to diagnose children who were behind their age cohort, so long as the information was used to provide specialized interventions so that the child could begin to make appropriate progress. Both questioned the wisdom of comparing children to their peers as opposed to an absolute criterion that could be used for diagnosing educational needs.

Dewey (1922/ 2008a) also argued against using the tests to place children in grades. Such homogeneously graded classrooms would treat the children in these classes as averages, or as he wrote, “mediocrities,” (Dewey, 1922/ 2008a, p. 291). Homogeneous groupings based upon intelligence test scores would fail to provide shared educative experiences for the different groups of children living together in a democracy (Dewey, 1922/ 2008a). Grouping children by intelligence would obscure the unique talents and interests of the individual child, instead treating children as members of superior, average, or mediocre intelligence class (Dewey, 1922/ 2008a, 1922/ 2008b).

Promoters of intelligence testing were much more interested in educating children with high IQs than those with average or low IQs. After all, in the world view of the promoters of intelligence tests, the future of the democracy rested with the proper education of gifted children

(Goddard, 1922; Terman, 1919, 1920; Whipple, 1922). Terman (1919, 1922a) in particular was a strong advocate for expanding the educational opportunities for gifted students, frequently writing of the "educational neglect of superior children," (1919, p. 165). In heterogeneous classes, the teacher had to focus time and attention to the behavioral problems caused by the "defectives," leaving the gifted children to languish. Civilization depended on the preparation of the gifted for future leadership and innovation (Terman, 1919). Therefore, educational opportunities designed to optimize the education of these children must be greatly expanded (Terman, 1920).

Bagley (1922b; 1925) took issue with the position that the gifted were neglected. He suggested that high schools and colleges already provided for the needs of the gifted (Bagley, 1925). Indeed, he repeated Galton's assertion that the gifted were capable of taking care of themselves (1922a). Rather, it was the average and below average in intelligence that needed additional support. Bagley (1922b) warned that the tests would "render a gratuitous and disastrous disservice if they encourage in the teacher the conviction that the illumination of common minds is either an impossible or a relatively unimportant task," (pp. 384-385).

Not only did promoters of intelligence testing argue for separate academic tracks based upon ability, but these tracks should be designed to prepare individuals for the kind of work for which they had the capacity (Terman, 1919; Goddard, 1922). Testing advocates believed that intelligence tests given at an early age were reasonably accurate predictors of the future limitations of a child's ultimate intellectual and, therefore educational, limitations. As such, tests could be used to tentatively place students in an academic track as early as five or six years of age. Indeed, Terman (1920) claimed that by the age of 12, the tests could accurately predict the upper limits of a child's intellectual capacity. Therefore, by the fifth year of school, students

could be tested and the scores used to guide them into vocations that were within the range of their predicted intellectual capacity.

The psychologists who advocated for intelligence testing were not seeking to choose an occupation for each child that they tested. Rather, they sought to identify a range, or cluster of occupations that were within a given child's capabilities (Terman, 1919). At this point, the child's "natural interests and practical considerations" would guide the child in choosing a vocation (Haggerty, Terman, Thorndike, Whipple, and Yerkes 1920; Terman, 1919). Therefore, Terman (1922a) did not view these recommendations as "deterministic," since the child did have some choice over the vocation in which he or she eventually was placed.

However, those who advocated for intelligence testing would have this choice limited to only those vocations that were within the predicted capacity of the child. For some children, this predicted capacity would place higher education beyond their reach. However, both Terman (1919) and Goddard (1922) believed that individuals with lower intelligence should be trained to do useful work. Therefore, a vocational curriculum should be provided for the intellectually inferior as an alternative to a curriculum that prepared students for entry into the university.

Indeed, modern industrialization and the corresponding rise in factories created a need for unintelligent workers to do monotonous work. For testing advocates, this gave even these inferior people a valued place in modern society. This dehumanizing view of people with low IQ scores as well as of factory workers was expressed by Terman (1919) when he wrote "The evolution of modern industrial organization together with the mechanization of processes by machinery is making possible a larger and larger utilization of inferior mentality. One man with ability to think and plan guides the labor of ten or twenty laborers, who do what they are told to

do and have little need for resourcefulness or initiative," (p. 276). Thus, testing could be used to find a use for even inferior people.

Both Lippmann (1922f) and Bagley (1922a, 1922b) were wary of power that testing for vocational placement and placement into the corresponding educational track potentially gave to the psychologists. Bagley was critical of the belief that the limits of a child's educability could be determined by the fifth or sixth year of schooling. He was also critical of the proposal to use the tests to determine whether a student received a vocational or intellectual education. He called the proposal "deterministic," in that the test would be used to limit individual opportunities.

Bagley (1922a, 1922b) and Lippmann (1922c) both noted that this deterministic use of the tests was particularly dangerous, given the high stakes involved, and the many factors other than intelligence that affected IQ score. Such factors included problems with the tests (Lippmann, 1922b), concerns over the meaning of test scores (Bagley, 1922b), the lack of a common definition of intelligence (Lippmann, 1922a, 1922b, 1922f; Bagley, 1922a), error (Lippmann, 1922b, 1923), lack of motivation, and poor testing taking ability (Lippmann, 1922c).

Lippmann claimed that the psychologists' proposal to use the tests to determine the future educational and career opportunities of children demonstrated their "will to power," (Lippmann, 1922f, p. 9). Lippmann warned that if the psychologists could make good on their claim that the tests could classify and sort children into different educational and vocational tracks, it would elevate them to "a position of power which no intellectual has held since the collapse of theocracy," (Lippmann, 1922f, p. 10).

The belief that teachers could do little to impact the intellect of students was also partly what Bagley meant when he labeled testing advocates "determinists." Bagley (1922a) described determinism as the belief that education and experience had little impact on an individual's

intelligence, and therefore the tests could be used to predict a child's future. Bagley (1922b) argued that this belief that intelligence was innate and immutable was based on hypotheses and assumptions. Furthermore, the use of the tests to make fatalistic assumptions about children's futures should be condemned (Bagley, 1922b).

Bagley argued that just because there may be inherent limitations to a child's capacity for abstract thought did not mean that there were limitations to what a child could learn. He noted that while not everyone possessed the intellectual capacity to discover abstract concepts such as those discovered by Newton, the average individual possessed the intellectual capacity to learn and understand these ideas. It was this ability of the individual of common intellectual capacity to learn and understand what great thinkers had discovered that allowed these discoveries to become part of the civilizations collective intelligence (Bagley, 1922a).

Responding to Bagley's (1922a) concern that there was not enough evidence in support of the test's ability to predict a student's upper limits, Terman (1922a) suggested that Bagley was going against the preponderance of evidence in insisting that intelligence testers prove a "universal negative," (Terman, 1922a, p. 60). "No one has ever proved the impossibility of life returning forty-eight hours after death, yet we do not hesitate to bury our dead within that time," (Terman, 1922a, p. 60). Bagley responded, "Yes, truly; yet we do not bury them before they die, although we may be very certain that death is imminent," (Bagley, 1922b, p. 372).

Lippmann (Lippmann, 1922d) was also very concerned over the proposal that test scores would be used to label and limit life prospects for children. In one of the most eloquent passages in the debates, he wrote, "If, for example, the impression takes root that these tests really measure intelligence, that they constitute a sort of last judgment on the child's capacity, that they reveal 'scientifically' his predestined ability, then it would be a thousand times better if all the

intelligence testers and all their questionnaires were sunk without warning in the Sargasso Sea. One has only to read around in the literature of the subject, but more especially in the work of popularizers like McDougall and Stoddard, to see how easily the intelligence test can be turned into an engine of cruelty, how easily in the hands of blundering or prejudiced men it could turn into a method of stamping a permanent sense of inferiority upon the seal of a child," (Lippmann, 1922d, p. 297).

Dewey (1922/ 2008b) was also concerned over the proposals to use the tests to classify and sort students into vocational tracks. In his view, such a use of tests would dramatically repurpose schools. To Dewey, the mission of the school was to release the abilities that resided within each child. The tests, however, would classify and sort students into a limited number of vocational options that were based upon the child's IQ scores, and predetermined by the needs of business. In this way, tests would create a new social caste system.

The assumption that school success or failure was caused by innate intelligence resulted in a shift of accountability for student outcomes from teachers and schools to the innate ability of students (Terman, 1919, 1920). Psychologists such as Terman believed that education had little impact on intelligence. Furthermore, both Goddard (1922) and Terman believed that schools and teachers could do little to help unintelligent children stay on grade level. Terman (1919) suggested that it was unfair to hold different teachers accountable for the same educational outcomes when they had different "material" (meaning students) with which to work.

Lippmann (1922c) and Bagley (1922a) both expressed concern about this view of the school's ability to affect the learning of students. Lippmann argued that if the school's function was merely to predict which students would succeed and which would fail, then the school was effectively impotent.

Bagley (1922a) claimed that if the testing advocates were right, then the teacher was merely a certifying agent, whose sole purpose was to indicate whether or not a particular child had the intelligence to move on. Rather, Bagley argued that he had witnessed "skillful and devoted teachers" help even dull children to learn.

Indeed, Bagley (1922a; 1925), Lippmann (1922c), and Bond (1934/ 1966) found it unethical to absolve schools from the responsibility of educating students. Neither was it ethical to condemn children to limited educational and occupational opportunities based on a single test score. As long as there was any possibility that these intellectual limitations might be based more on the assumptions of intelligence testers than on facts, educators had an obligation to continue to try to provide all students with an intellectual education.

Bond (1924b, 1934/ 1966) noted that the low IQ scores of Black and poor Southern white children were likely caused by factors in the environment, including the quality of schooling to which these children had access. Further limiting educational and occupational opportunities for these children based upon their IQ scores was unethical. To the contrary, the low IQ scores of these populations was reason to improve the educational opportunities for these children.

Bond (1924b) suggested that the tests could function to standardize schools and the curriculum. As such, they could become a means to provide equity in education. In this way, he anticipated the standards and accountability movement's argument that tests could guarantee that all students had equal access to the same curriculum.

Democracy

Finally, the belief that intelligence tests identified the upper limits of a child's capacity to learn had implications for democracy. Indeed, at least three competing definitions of democracy

came out of the debates. Terman (1922a), Goddard (1922) and Whipple (1922) believed that the "real meaning of democracy" was "equity of opportunity," (Whipple, 1922, p. 602). They believed that democracy depended upon leadership by the best. Leaders should be chosen based solely on merit, and merit was defined as inherent intelligence. Accordingly, "equity of opportunity" (Whipple, 1922, p. 602) meant that any inherently talented individual had the opportunity to rise to positions of leadership regardless of "race," gender, or station of birth. To be sure, Terman (1919) suggested that lack of occupations open to women caused a waste of intellect that he found appalling. This was a liberal perspective, given the time period.

According to this perspective of democracy, innate intelligence was not merely the criteria for entry into leadership positions; it also could be a criterion for voting. "What about democracy," wrote Goddard (1922), "can we hope to have a successful democracy where the average mentality is thirteen?" (Goddard, 1922, p. 96). Goddard was particularly concerned that decision making should not be placed in the hands of unintelligent people. He went so far as to argue that the unintelligent should be disenfranchised, stating that "While we all believe in democracy, we may nevertheless admit that we have been too free with the franchise and it would seem a self-evident fact that the feeble-minded should not be allowed to take part in civic affairs; should not be allowed to vote. It goes without saying that they cannot vote intelligently, they are so easily led that they constitute the venial vote and one imbecile who knows nothing of civic matters can annul the vote of the most intelligent citizen," (1922, p. 99).

This role of intelligence tests as gatekeepers for both leadership and participation troubled Bagley (1923). Such a use of the tests created winners and losers, leaders and followers, intellectuals and menial workers. This was anathema to a democracy, where the goal should be "…the absolute advancement of both one and all," (p. 41).

Bagley (1922a) suggested that a true democracy elevated "the common man to a position of supreme collective control," (Bagley, 1922a, p. 380). This democratic ideal required that ideas be broadly disseminated in order to promote informed collective judgment. Democracy depended upon an educated citizenry that enabled "the common man to choose his leaders wisely, scrutinize their programs with sagacity and, in the pungent slang of the day, tell them 'where to get off' when they go wrong," (Bagley, 1922a, p. 380). Thus, in a democracy, "…the education of the great masses of the people is of vastly more significance to any nation than is the refined and advanced training of the few,—important as I gladly admit the latter to be," (Bagley, 1925, p. 36).

Democratic institutions also depended on the dissemination of a common culture and values (Bagley, 1923; 1925; Dewey, 1922a). Bagley was concerned that the testing and sorting being proposed by the testing advocates would not only disenfranchise a large portion of the population from the life of the mind, but would also prevent the cultivation of ideas, of shared values, and of a common culture that was necessary for democracy to thrive. Bagley (1922a, 1925), and Dewey (1922/ 2008a) also argued that education in a democracy provided a diverse populace with a common *experience* that would enable them to understand each other, to live together and to work toward a common purpose. Bagley (1925) argued that this common educational experience made people more alike, which acted as a social leveler.

Finally, in addition to a concern that the testing advocates sought to exclude a large portion of the population from the life of the mind, Bagley (1922a) also expressed his concern that intelligence was the sole criteria by which testing advocates sought to choose future leaders. He suggested that other qualities may be just as important as intelligence, and that possession of these other qualities could make up for average or even below average intelligence. Such

qualities might include ‘“human qualities’, such as sympathy, tact, humor, and sociability, and 'moral' qualities, such as integrity, industry, persistence, courage, and loyalty,” (1922a, p. 380-381). For Bagley, a leader must be more than a thinking machine.

Dewey had a somewhat different take on democracy. For Dewey (1922/ 2008b), democracy was not about rule by the masses or universal suffrage. Nor did he accept democracy as freedom to achieve success based upon personal merit. Rather, “The democrat with his faith in moral equality is the representative of aristocracy made universal. His equality is that of distinction made universal,” (1922/ 2008b, p. 300 of *John Dewey, the Middle Works*). Therefore, in contrast to intelligence testers who held that people were inherently different and democracy meant identifying the merit among the masses, Dewey argued that a democracy was composed of unique individuals, with unique talents and interests, but with equality of status in spite of these differences. Furthermore, democracy allowed each individual to fully develop his or her unique abilities.

While it could be argued that one of the goals of the intelligence testers was to individualize education according to the abilities of the individual, Dewey did not see it this way. He argued that testing and sorting people into groups based on a limited number of tested measures submerged their individuality. They became classes of people: “A Men,” superiors, inferiors, leaders, and menial workers. Even within these groups, they became averages, or, as Dewey put it, “mediocrities,” (Dewey, 1922/ 2008a, p. 291).

In addition, Dewey (2008/1922b) expressed concern that IQ classifications were always relative to other people. Dewey pointed out that superiority or inferiority of any given trait did not equate to superiority or inferiority as a human being, and should not be treated as such. Different cultures placed more or less value on different traits. Indeed, different individuals

placed more or less value on different traits based upon uniquely personal values, interests, talents, and goals.

Conclusion

The search for a method to objectively measure the quality of a human being led to the development of the intelligence test. The correlation of scores on these tests to other measures of social and economic success reinforced the belief that the tests measured human quality, and led proposals for using the tests to make schools and society more efficient. In addition, these beliefs that the tests measured human quality, and the correlation between test scores and social and economic success also reinforced and rationalized beliefs in a biologically determined class and "racial" hierarchy.

The claims and proposals made by the intelligence testers, and quality of evidence with which they backed these claims, led critics to counter these claims and proposals.
The resulting debates tended to center around several emerging themes. These included the following: intelligence as merit; the varied definitions of democracy; the validity of the tests as measures of intelligence; tests as both a means of creating opportunities for historically marginalized groups and as a rationale for the existing "race" and class based caste system; technological/ objective method for making school and society more efficient; and biological determinism. These themes will be compared, and if possible synthesized, with themes identified during the subsequent intelligence testing and accountability testing debates. These enduring themes and their implications for testing policy will be discussed in Chapter 6.

Chapter 4: The Resurgence of the Intelligence Testing Debates

The IQ testing debates resurfaced in 1969 with Arthur Jensen's article, *How Much Can We Boost IQ and Scholastic Achievement?* and again in 1994 with the publication of Richard Herrnstein and Charles Murray's *The Bell Curve: Intelligence and Class Structure in American Life*. Both works argued that human intelligence was largely hereditary and immutable (Jensen, 1969, pp. 17-19 29-88; Herrnstein & Murray, 1994, pp. 23; 105-110). Herrnstein and Murray claimed that that many societal problems were caused by less intelligent people (Herrnstein & Murray, 1994, Chapters 5-12, 14-16), and therefore money spent to meliorate social problems could have little or no lasting effect (Herrnstein & Murray, 1994, Chapters 8, 9, 17, 18; Jensen, 1969, pp. 2-5, 104-108). In addition, both works concluded that the variations in mean IQ test scores of different "race" and ethnic groups were primarily genetic in origin (Herrnstein & Murray, 1994, pp. 300-315; Jensen, 1969, pp. 71-72, 79-88), and therefore little could be done to close ability and achievement gaps between minorities and non-minorities (Herrnstein & Murray, 1994, chapter 17; Jensen, 1969, pp. 2-5, 104-108), although Jensen suggested that compensatory programs that focused on associative learning could close the scholastic achievement gap (Jensen, 1969, pp. 59,111-117). Both works dismissed the possibility that endemic societal problems such as racism or social and economic inequalities created barriers to advancement for minority groups. Rather, they placed the blame for the lack of advancement of poor and minority people on traits supposedly inherent to members of these groups (Apple, 2001).

Jensen (1969) and Herrnstein and Murray (1994) exhumed arguments that had "been discredited intellectually many times before," (Apple, 2001, p. 52). Both works interpreted the

available data from a hereditarian and racialist perspective, with a primary objective of supporting arguments for ending compensatory education, welfare, and affirmative action.

Following the publication of "How Much Can We Boost IQ and Scholastic Achievement" (Jensen, 1969) and again after the publication of *The Bell Curve* (Herrnstein & Murray, 1994), a number of scholars from varied disciplines responded quickly to counter the arguments put forward by these works. This chapter follows a slightly different format than Chapters 3 or 5, in that the advocates and critics of testing will be discussed separately, instead of alternately. The section covering Jensen (1969) and Herrnstein and Murray (1994) is organized by the following topics: The nature of intelligence; intelligence and social problems; "race;" and compensatory education and other social welfare programs. The section covering the response of the critics will cover the following topics: political agenda; unsound methodology; the nature of intelligence; understanding of modern genetic theory; "race;" validity of tests; and compensatory education. The analysis of the arguments in this section will be used to identify emerging themes regarding testing policy in the United States.

Argument of the Intelligence Testers

The Nature of Intelligence

As in the earlier IQ testing debates from the 1920s, finding a common definition of the construct of intelligence was still problematic in the late 1960s. Suggesting that measuring a thing could help flesh out the construct, Jensen (1969, p. 8) repeated Boring's (June 6, 1923) definition that intelligence was what intelligence tests measured. However, Jensen also provided a working definition of intelligence as the "capacity for abstract reasoning and problem solving,"

(p. 19). In addition, he repeated O. D. Duncan's assertion that intelligence was a socially defined construct that either consciously or unconsciously incorporated traits that were directly related to success in modern industrialized societies. Indeed, it was no accident that there was a high correlation between IQ scores and occupational prestige (O.D. Duncan, 1968, pp. 90-91). At the same time, Jensen believed that this socially defined construct was a real phenomenon that had its basis in biology (Jensen 1969, pp. 19-20).

Both Jensen (1969, p. 9) and Herrnstein and Murray (1994) took pains to demonstrate that both the existence of general intelligence (*g*) as well as the ability of the tests to measure *g* had been established beyond doubt (Gould, 1981/ 1996; Herrnstein & Murray, 1994, pp. 6-27). While noting that *g* was "a hypothetical construct intended to explain covariation among tests," Jensen also suggested that *g* had "stood like a rock of Gibraltar in psychometrics, defying any attempt to construct a test of complex problem solving which excludes it," (Jensen, 1969, p. 9). In addition, Jensen claimed that *g* was polygenic, meaning that it was the sum effect of a number of different genetic traits. For Jensen, this polygenic trait was synonymous with intelligence.

Like Jensen (1969, p. 9-10), Herrnstein and Murray (1994) also argued that *g* was a real construct that was measured by intelligence tests. They claimed that *g* had been established beyond all doubt, and referred to individuals such as Howard Gardner and Robert Sternberg, both of whom proposed alternatives to the theory of *g*, as "radicals." Like Jensen, Herrnstein and Murray (1994) suggested that intelligence was synonymous to whatever people mean when discussing IQ scores. Finally, they claimed that all academic tests measured *g*, but IQ tests were designed to measure *g*, and were therefore much more accurate.

Both Jensen (1969) and Herrnstein and Murray's (1994) arguments depended upon intelligence being primarily hereditary and immutable. An IQ score was a phenotype, the

outward manifestation of a trait resulting from the interaction between the genotype and environmental factors. Jensen used the so called heritability coefficient to parse out the variance in IQ scores that was caused by variations in genotype from that portion caused by variations in the environment.

The heritability coefficient was the square of the correlation between genotype and phenotype (Jensen, 1969, p. 17). By definition, heritability was greater in homogeneous environments, and smaller in heterogeneous environments (Jensen, 1969 pp. 43, 45). This was because environment affected phenotypes but not genotypes. Where all members of population experienced very similar environmental conditions, even where those conditions were poor, environment would have a relatively uniform effect on phenotypes across individuals in the population. Under such conditions, variations in the phenotype could be attributed to the genes. Therefore, a stable environment increased the correlation between genotype and phenotype, and heritability would decrease.

Conversely, where members of the population experienced very different environmental conditions, environment was more likely to affect phenotype, making it less likely that variations in phenotype were genetic in origin. Therefore, an unstable environment decreased the correlation between genotype and phenotype, and heritability would increase. As such, as both Jensen (1969, p. 43) and Herrnstein and Murray (1994, pp. 105-110) noted, the heritability coefficient could not be considered to be a constant, applicable to any population. It was specific to the population studied and the range of environmental conditions that this population experienced at the time they were studied (Jensen, 1969, pp. 42-43).

Heritability estimates used by both Jensen (1969 pp. 46-54) and Herrnstein and Murray (1994, pp. 107-108) were based upon calculations from studies that attempted to control for

genetic and environmental variability. Genetic variability could be controlled by studying people with identical genotypes, as was the case in monozygotic twins. Environmental variability could be somewhat controlled in studies where siblings were raised in the same and different households. In the case of monozygotic twins, all phenotypic variation must be due to differences in the environment. Therefore, researchers could isolate the effects of the environment by comparing variations in IQ scores of monozygotic twins reared in different households. Based on the higher correlation of intelligence test scores among monozygotic twins reared apart compared to those reared separately or dizygotic twins reared separately, Jensen, and Herrnstein and Murray concluded that intelligence was highly heritable.

Despite the fact that both Herrnstein and Murray (1994, pp. 105-110) and Jensen (1969, pp. 42-43) acknowledged that heritability varied with environment and was therefore a population statistic, and even a relatively high heritability coefficient still left a good deal of the phenotype to be explained by the environment (Herrnstein & Murray, 1994, p. 109), both Jensen (See, for example pp. 78-88, 95) and Herrnstein and Murray (1994, chapters 13, 17-20) treated heritability coefficients as though they were close to 1.00, and both applied heritability coefficients calculated on one population to another population. Both dismissed arguments that poor and minority populations experienced mean environmental conditions that were much different than Caucasian monozygotic twins reared apart. As such, both concluded that the variations in mean IQ scores between different racial, ethnic, and social class groups were caused by variations in the genotype endemic to these populations. Therefore, no amount of manipulation of the environment would significantly improve intelligence or achievement.

Intelligence and Social Problems

Because of the relationship between the stability of the environment and the heritability coefficient, Herrnstein and Murray (1994, pp. 91 and 106) argued that efforts to provide equal environments for all children were futile because such efforts only increased the heritability of IQ. This was a true, though somewhat meaningless statement. Increasing the uniformity of environmental experiences would increase the confidence with which variations in IQ scores could be attributed to variations in genotype. However, if environmental experiences varied greatly to begin with, then it is likely that decreasing environmental variation would narrow the spread of IQ and achievement test scores, likely improving scores for individuals at the lower end of the spectrum.

In making their case against compensatory education and other liberal social policies aimed at meliorating social problems, Herrnstein and Murray (1994, Part II, pp. 117-266) set out to demonstrate that social problems were caused by genetically unintelligent people as opposed to social or environmental inequities. They performed a series of simple factor analyses comparing the impact of SES to that of IQ on a variety of social problems. In this way, they demonstrated that IQ scores explained more of the variations in the occurrence of the social problem in question than did SES. They concluded that low innate intelligence caused poverty, school retention and drop out, worker accidents and disability, unemployment, out of wedlock births, divorce, general immorality, low birth-weight babies, infant mortality, child abuse and neglect, criminality, low civic mindedness, poor quality of home environment, poor parenting skills, and low ratings on the "middle class values index."

At times, Herrnstein and Murray (1994) also used syllogisms to make their point that IQ caused social problems. For example, they determined that intelligence caused child abuse and

neglect using the following logic: there was a correlation between SES and child abuse and neglect; IQ was related to or caused SES; therefore IQ caused child abuse and neglect (Herrnstein & Murray, 1994, pp. 207-213).

Race

To give the impression that they were unbiased against minority "races," Herrnstein and Murray (1994, p. 125) limited their discussion of the relationship between IQ and various social problems to white subjects. However, their agenda, as well as that of Jensen (1969), involved discrediting public policies aimed at redressing problems of past discrimination. It was therefore necessary to show that such policies were futile by attributing problems experienced by minority "races" to their inherent inferiority.

After the authors of both works made the point that people should be judged on their individual merit rather than as members of a group (Herrnstein & Murray1994, p. 450; Jensen, 1969, pp. 78-79), both then cited a number of studies that seemed to indicate that "races" could be ranked according to mean IQ scores, with Blacks consistently scoring at the bottom (Herrnstein & Murray, 1994, Chapter 13; Jensen, 1969, pp. 78-88). Mean Black IQ scores have consistently fallen one full standard deviation below the mean for whites, and gaps between these means have persisted at all SES levels. While Herrnstein and Murray (1994) admitted that controlling for SES explained as much as 37 percent of the Black/ white IQ score gap, they attributed at least some of this to the effects of IQ in determining SES, rather than the effects of SES on IQ. Indeed, both Jensen (1969, p. 75) and Herrnstein and Murray (1994, see for example p. 305) argued that IQ caused a large portion of SES, rather than the other way round. This was

yet another way that the authors of both works dismissed the possible effects of the environment on explaining these differences.

Even if environmental factors explained some of the variations between mean Black and White IQ scores, Herrnstein and Murray (1994, p. 298-299) argued that it was impossible for the environment of Blacks to differ so dramatically from whites as to produce such large differences. They claimed that the mean environment for Blacks would have to be 1.58 standard deviations worse for Blacks than for whites. This argument was based on a 60% heritability estimate, the assumption that all important environmental variables that made up the environment were known and could be accurately measured, and that the environmental variation for the population of Blacks was the same as that for the population on which their estimated heritability coefficient was calculated.

Even if mean IQ differences between Blacks and whites were caused by significantly worse environmental exposures experienced by Blacks, Herrnstein and Murray (1994, pp. 314-315) claimed that there was little that could be done to narrow these gaps. Such environmental factors were too profound, and social policies created to meliorate them would be too expensive to have any real impact.

In addition to arguing that variations in mean IQ scores between Blacks and whites must be genetically based since there was a gap at all levels of SES, both Jensen (1969, p. 81) and Herrnstein and Murray (1994, pp. 280-286) also argued that the gap could not be due to cultural bias incorporated into the tests. Jensen noted that Blacks actually performed slightly worse compared to whites on so-called culture-fair tests than on conventional IQ tests. Alternatively, Herrnstein and Murray reasoned that the validity of any test was determined by its ability to predict something. Since IQ tests were good predictors of the performance of Blacks in college

and on the job, they concluded that IQ tests must be valid. Furthermore, they, too, noted that the gap between Blacks and whites was wider on items that appeared to be culturally neutral than on those items that appeared to rely on cultural knowledge. Therefore, they concluded, the IQ tests were not culturally biased against Blacks.

Herrnstein and Murray (1994, pp. 289-295) provided mixed evidence that the IQ gap between Blacks and whites was shrinking, although their own data from the NLSY suggested that it was growing. They also discussed the closing of the achievement gaps on the SAT, ACT and NAEP tests. They explained the closing of these gaps by suggesting that the environment for Blacks had improved following the civil rights movement. This included diminishing racism. Furthermore, schools had improved for Blacks during this period, but not for whites. Indeed, they argued, schools had actually grown worse for higher performing students. This supported their claim that schools were failing in general, and failing the gifted in particular. Here, "Black" was synonymous with low performer, and "white" with high performer. Therefore, they argued, SAT scores were improving for Blacks not because they were moving into the highest achievement levels, but because they were moving out of the lowest achievement levels.

Compensatory Education and Other Social Welfare Programs

Not only was it necessary for the authors of the respective works to show that factors inherent to membership in certain economic and racial groups caused the failure of these groups to advance, they also needed to show that interventions, in general, were ineffective. To demonstrate this, they looked at a number of such programs. Those social programs that were most directly related to education will be examined here. These included the role of welfare, compensatory education, education for the gifted and affirmative action in higher education.

Both *How Much Can We Boost IQ and Scholastic Achievement* (Jensen, 1969) and *The Bell Curve* (Herrnstein & Murray, 1994) attempted to make a case that welfare had little impact on the cognitive ability of the poor, and actually exacerbated existing social problems. Jensen (1969, pp. 73-74), for example, after citing evidence that nutritional supplements could have profound effects on cognitive function, then claimed that nutrition was a threshold variable and that malnourishment below this threshold was rare in the United States. Therefore, according to Jensen, welfare in the form of food assistance would have little impact on the IQ scores of poor children.

Furthermore, Jensen (1969, p. 95) claimed that welfare policies aimed at meliorating the effects of poverty among African Americans without "eugenic foresight" could exacerbate dysgenic effects already at work, leading to "the genetic enslavement of a substantial portion of our population," (Jensen, 1969, p. 95). The section implies that welfare would drive down the collective intelligence of African Americans by allowing greater numbers of poor, low IQ African American's to survive.

Herrnstein and Murray (1994, Chapter 8 and 15) went even farther, suggesting that welfare policies provided financial incentives for poor, unintelligent women to have more children. This would have a dysgenic effect on the population as a whole.

Indeed, both Jensen (1969, p. 95) and Herrnstein and Murray (1994, Chapter 8 and 15) suggested that certain segments of the low IQ population had more children and with less time between subsequent generations than more intelligent segments of the population. Although Herrnstein and Murray claimed that these dysgenic effects were in part caused by welfare, and therefore social class, both Jensen (1969) and Herrnstein and Murray (1994) claimed that the reproductive habits of certain minority groups exacerbated these dysgenic effects.

Nowhere in their arguments of the reproductive habits of intelligent and unintelligent individuals did Herrnstein and Murray (1994) ever discuss the responsibility that men had in reproduction or raising children. Nor did they acknowledge the possibility that there might be other motivations for engaging in reproductive acts than to make money off of welfare.

Herrnstein and Murray (1994 Chapter 15) claimed that welfare exacerbated dysgenic pressures in other ways, as well. They claimed that welfare created an incentive for poor immigrants with low cognitive abilities to immigrate to the United States, since they were more prone to economic and occupational failures than their more intelligent peers, and because there was no such social safety net in their respective home countries. They arrived at this conclusion by calculating the proportion of immigrants to the United States that belonged to different ethnicities, and assigning each ethnicity mean IQ score based on the work of Richard Lynn. They did not use actual IQ scores. Nor did they provide evidence to support their claim that lower IQ immigrants were coming to the United States for the social safety net. Finally, they did not consider the many reasons that immigrants came to the United States, including but not limited to, escape from persecution, to pursue employment opportunities, or gain an education.

Both Jensen (1969, p. 85-86, 96-109) and Herrnstein and Murray (1994, Chapter 17) also argued that attempts to meliorate problems associated with poverty and racism through compensatory education programs were futile. Both *How Much Can We Boost IQ and Scholastic Achievement* (Jensen, 1969, 85-86) and *The Bell Curve* (Herrnstein & Murray, 1994, pp. 394-396) used the Coleman Report (Coleman, et al., 1966) to support their claim that increased funding for schools would have little impact on the target populations. Coleman, et al. (1966) found that environmental factors affected achievement and ability within groups, but this finding did not explain between group variations. These findings were interpreted as indicating that

attempts to equalize schooling had hit a threshold in which additional resources would have little effect (Herrnstein & Murray, 1994, p. 396) and that ability and achievement gaps between groups would be unaffected by environmental interventions (Jensen, 1969, pp. 85-86).

Furthermore, both Jensen (1969) and Herrnstein and Murray (1994) argued that compensatory programs would only have an impact for children from severely deprived environments. Jensen (1969, p. 90) argued that, like nutrition, environment was a threshold variable. By this he meant that modifying the environment would only have an impact on intelligence if the quality of the environment was below a certain minimum threshold. Above this threshold, such manipulations would have little effect.

Likewise, Herrnstein and Murray (1994, p. 390) claimed that only a small proportion of children lived in an environment so grim that changing the environment would have an impact on IQ. Indeed, they claimed that the only intervention that would likely affect IQ in young children was adoption at birth from a bad environment to a good one (Herrnstein & Murray, 1994, p. 410). To be sure, they argued, the United States had done everything possible to narrow the IQ gap between groups when they instituted compulsory kindergarten through 12 grade education (Herrnstein & Murray, 1994, p. 414). Therefore, they dismissed the possibility that additional funding for compensatory education would have any measureable effect.

Indeed, both Jensen (1969) and Herrnstein and Murray (1994, pp. 402-410, 415) argued that programs such as Head Start had little lasting effect on IQ. Jensen suggested that any increases in IQ observed in children participating in compensatory programs such as Head Start were not "real." He suggested that such increases might be due to regression toward the mean, improvements in lower level thinking skills (such as memorization), learning the test, or "the hothouse effect," which he described as analogous to forcing bulbs to bloom by placing them in

a hot house. In the latter instance, increasing stimulation for very young children could stimulate superficial, short-lived improvements in intelligence test scores. Jensen dismissed a study that found over one full standard deviation improvement in IQ resulting from a compensatory education program by suggesting that the study needed to be replicated.

Similarly, Herrnstein and Murray (1994, Chapter 17) dismissed the small IQ gains that studies showed had resulted from some compensatory programs. Furthermore, they argued that these gains were subject to "the dreaded fade out effect," meaning that any gains faded once the child was no longer in the program (Herrnstein & Murray, 1994, p. 403). They failed to consider that children typically returned to the deprived environment once the compensatory program had ended, choosing instead to suggest that any gains resulting from compensatory programs were transient, at best. Finally, they argued that those studies that did show significant gains in IQ points were either methodologically flawed, or were very intense and therefore too expensive.

Although Jensen (1969) maintained that compensatory programs probably had little impact on IQ, he hypothesized that education targeted at the strengths of low IQ children could improve scholastic ability (Jensen, 1969, pp. 11-117). Jensen presented evidence that children in some minority groups were better at "associative learning" than at higher level cognitive skills. He suggested that interventions that played to the strengths of the target groups could close scholastic achievement gaps.

Rather than waste money on what they considered to be futile attempts to improve IQ or achievement for disadvantaged children, Herrnstein and Murray (1994, chapter 18) suggested that at least a portion of these resources would be better spent helping the gifted achieve their full potential. In arguments reminiscent of those previously made by Terman and Goddard, Herrnstein and Murray argued that the preparation of the gifted to assume their proper place in

society had originally been one of the primary goals of education. “It needs to be said openly: The people who run the United States—create its jobs, expand its technologies, cure its sick, teach in its universities, administer its cultural and political and legal institutions—are drawn mainly from a thin layer of cognitive ability at the top,” (Herrnstein & Murray, 1994, p. 418).

Indeed, using much of the same data cited by the critics of American education during the 1980s and 1990s, Herrnstein and Murray (1994, chapter 18) concluded that overall, education in the United States was improving. However, there was one exception to this improvement: public education was failing to properly educate the gifted. To be sure, they claimed, social policies of the 1960s and 1970s which focused on improving education for students with disabilities and poor and minority students had the effect of lowering educational standards. Rather than providing resources to the gifted, they argued, American educational policy shifted resources to low income students with learning problems.

Herrnstein and Murray (1994) went so far as to suggest that children in the lowest quartile of the IQ distribution were not only expendable; they were a net drag on society. Providing resources to educate those with low IQs was a waste of money. “…all the fine rhetoric about ‘investing in human capital’ to ‘make America competitive in the twenty-first century’ is not going to be able to overturn this reality: For many people, there is nothing they can learn that will repay the cost of the teaching,” (Herrnstein & Murray, 1994, p. 520).

As an alternative to compensatory education, Herrnstein and Murray (1994, pp. 440-445) provided a number of policy recommendations. These included giving parents, including low income parents, school choice (Herrnstein & Murray, 1994, pp. 440-441). They recommended redirecting at least some of the resources spent on poor, minority and disabled children to funding for programs for the gifted (Herrnstein & Murray, 1994, p. 441-442). In addition, they

recommended providing scholarships to gifted children from all socioeconomic backgrounds (Herrnstein & Murray, 1994, pp. 441-442). Finally, even as they noted that schools were segregated based on the wealth of the community, they warned that efforts to forcibly desegregate schools by "race" or class was "a cure far deadlier than the disease," (Herrnstein & Murray, 1994, p. 443).

Not only did Herrnstein and Murray (1994, Chapter 19) believe that welfare and compensatory education were a waste of resources, they also claimed that attempts to redress past discrimination by creating a strong, self-sustaining minority middle class through affirmative action programs undermined the United States meritocracy and was discriminatory against more highly qualified whites. While affirmative action programs were aimed at addressing the effects of discrimination in both the workplace and in higher education, this analysis will examine the arguments specific to affirmative action in higher education.

For Jensen (1969, p. 76), as well as Herrnstein and Murray (1994) the United States was a meritocracy, merit was genetically determined, and this merit could be measured by intelligence tests. Indeed, they argued, the traits measured by intelligence tests were particularly important in determining success in higher education. Furthermore, Herrnstein and Murray argued that if IQ was a measure of inherent capacity for success in higher education, then allowing an individual with a lower IQ to displace an individual with a higher IQ solely on the basis of "race" was discriminatory.

Herrnstein and Murray (1994, p. 449) claimed that affirmative action policies were based on the assumption that there were no inherent differences between "races". This was at least part of the reason that they took such pains to demonstrate that racial differences in IQ were both real and inherent. Affirmative action policies were harder to justify if population differences in mean

IQ were due to natural variations in genetic material instead of environmental variables related to discriminatory treatment or poverty.

Indeed, when researchers controlled for IQ scores, Herrnstein and Murray (1994, Chapter 19) claimed that Black and Hispanic students were actually overrepresented in universities. Not only was this unfair to the otherwise qualified white students, they argued, it set Black and Hispanic students up for failure. Higher failure rates harmed the self-esteem of individual Black and Hispanic students. Furthermore, it harmed their white classmates by promoting a negative perception of minority students. According to Herrnstein and Murray (1994), these negative perceptions were the cause of racial conflicts on college campuses. Therefore, affirmative action was bad for majority and minority students, alike.

As discussed above, both Jensen (1969, p. 78) and Herrnstein and Murray (1994, p. 450) argued that people should be judged according to their individual merit. The authors of both works were arguing against policies that gave advantages to people based solely on their affiliation with specific racial, ethnic, or economic class groups. For these authors, equality of opportunity should be based solely on an objective measure of merit, and merit was defined as one's score on an IQ test.

To bolster minority college enrollment, Herrnstein and Murray (1994, pp. 448, 475-446) recommended that colleges actively recruit poor and minority students whose IQ qualified them to compete with white students for college spots, but only use "disadvantage" as a criteria when two otherwise equal candidates were being considered. They noted that this would result in fewer Black college students at Harvard and Yale, but that the cream of the Black crop could still go to lesser universities.

Critics Respond

Both "How Much Can We Boost IQ and Scholastic Achievement" and *The Bell Curve* sparked an immediate and strong response from the critics of these works. Because both works dealt with human limitations as well as "race," the response of several of the critics was emotionally charged, and at times fell into the trap of overreaching or reverting to innuendo to support their arguments (Heckman, 1995). As Heckman noted, many of the claims made by Herrnstein and Murray were empirically correct. These included the claims the tests were not culturally biased (Herrnstein & Murray, 1994, Chapter 13), demographic groups differed by intelligence test scores (Herrnstein & Murray, 1994, Chapter 13), and intelligence test scores predicted social and educational success (Herrnstein & Murray, 1994, Part I). Therefore, some of the arguments presented below were actually alternative explanations for the empirical evidence provided by Jensen (1969) and Herrnstein and Murray (1994). In this respect, they sometimes fell into some of the same traps, such as insinuating causation from a correlation.

The critiques of "How Much Can We Boost IQ and Scholastic Achievement" (Jensen, 1969) and *The Bell Curve* (Herrnstein & Murray, 1994) tended to take the following positions: the authors of both works relied on unsound scientific methods and conclusions to support a conservative political agenda (Gould, 1981/ 1996; Heckman, 1995; Kamin, 1999; Lane, 1975/ 1999; Rury, 1995; Ryan, 1999; Wallace, 1968); the existence of *g* was neither demonstrated nor necessary to support the claims made by Jensen (1969) or Herrnstein and Murray (1994) (Gould, 1981/ 1996; Heckman, 1995); both works demonstrated a lack of understanding of genetics (Gould, 1974/ 1975/ 1999, 1981/ 1996; Lewontin, 1975/ 1999; Montagu, 1975/ 1999a, 1975/

1999c); the heritability coefficient was of limited use and was misapplied by Jensen (1969) and Herrnstein and Murray (1994) (Biesheuvel, 1972/ 1975/ 1999; Bodmer, 1972/ 1975/ 1999; Brace & Livingston, 1972/ 1975/ 1999; Montagu, 1975/ 1999a; Bronfenbrenner, 1972/ 1975/ 1999; Gordon & Green, 1974/ 1975/ 1999; Gould, 1974/ 1975/ 1999, 1981/ 1996; Heckman, 1995; Sanday, 1972/ 1975/ 1999); "race" was a controversial construct (Lieberman, Littlefield & Reynolds, 1999; Montagu, 1975/ 1999a, 1975/ 1999c, 1997, 1941) that was socially defined and used to make biological inferences (Fried 1968; Gordon & Green, 1974/ 1975/ 1999; Lieberman, Littlefield & Reynolds, 1999; Montagu, 1975/ 1999c); intelligence tests measured something other than intelligence (Heckman, 1995); the tests were culturally biased (Biesheuvel, 1972/ 1975/ 1999; Gould, 1974/ 1975/ 1999, 1996/ 1981; Kagan, 1975/ 1999; Montagu, 1975/ 1999a; Sanday, 1972/ 1975/ 1999); research on compensatory programs showed they were more effective than Jensen (1969, pp. 96-109) or Herrnstein and Murray (1994, Chapter 17) portrayed them to be (Barnett, 1995; Bronfenbrenner, 1974/ 1975/ 1999; Lee, Brooks-Gunn, Schnur, Liaw, 1990; Ryan, 1999); neither Jensen (1969) nor Herrnstein and Murray (1994) conducted a cost benefit analysis of compensatory education interventions before dismissing their effectiveness (Heckman, 1995); there was a moral obligation to continue attempts to help all children to achieve their full potential (Bodmer, 1972/ 1975/ 1999; Bronfenbrenner, 1972/ 1975/ 1999; Gould, 1974/ 1975/ 1999; Lewontin, 1970/ 1975/ 1999; Luria, 1974/ 1975/ 1999); and finally, the conclusions and solutions suggested by Herrnstein and Murray (1994, Chapters 21 and 22), in particular, did not follow from the evidence that they provided (Heckman, 1995).

The remainder of this chapter is divided into sections by topic of the critics' responses. These topics include: political agenda; inappropriate methodology; the nature of intelligence; modern genetic theory; "race;" validity; and compensatory education.

Political Agenda

Several of the critics of "How Much Can We Boost IQ and Scholastic Achievement*"* and *The Bell Curve* argue that both were written to promote a specific agenda. Indeed, the first line of Jensen's 1969 article expressed the belief that "Compensatory education has been tried and it apparently has failed," (Jensen, 1969, p. 2; Lewontin, 1970/ 1975/ 1999). With this pronouncement, Jensen established his position that resources dedicated to early childhood compensatory education programs for poor and minority children were wasted. Herrnstein and Murray (Chapter 17 and 19) took a similar position, arguing that compensatory education and affirmative action programs were futile since IQ, the cause of the problems these programs were intended to meliorate, was primarily inherent and immutable.

Rury (1995) noted that although early IQ testing was embraced by many liberals who saw intelligence testing as a means to create avenues for social mobility that were based upon merit rather than birthright, contemporary arguments that intelligence was measureable, inherited, and immutable were associated with conservatives who lobbied for limited resources to be dedicated to those with the most ability. Because the distribution of IQ scores was related to "race" and class, such policies reinforced existing social class structures.

Similarly, Wallace (1968) argued that whereas revolutionary societies tended to place educational emphasis on intellectual and moral education, conservative and reactionary societies shifted educational emphasis to focus more on "technic" and moral development. This shift away from intellectual training was a defensive mechanism aimed at suppressing the dissemination of subversive ideas, whereas training in "technic" prepared individuals to serve utilitarian ends.

Such a shift was intended to maintain the existing social order (Wallace, 1968). Gould (1981/ 1996) suggested that this shift away from intellectual development during times of conservative stability or retrenchment was aligned with a belief in intelligence as inherent and immutable.

Likewise, Gould (1981/ 1996) suggested that conservatism was associated with a shift toward meritocracy and away from government intervention. He noted that "The resurgences of biological determinism correlate with episodes of political retrenchment, particularly with campaigns for reduced government spending on social programs, or at times of fear among ruling elites, when disadvantaged groups sow serious social unrest or even threaten to usurp power," (Gould, 1981/ 1996, p. 28). He went on to note that Jensen published "How Much Can We Boost IQ and Scholastic Achievement" (1969) just prior to the election of Richard Nixon, and the publication of *The Bell Curve* (1994) coincided with the Republican revolution of the mid-1990s.

Unsound Methodology

While Jensen (1969) and Herrnstein and Murray (1994) were fairly straightforward regarding their respective policy agendas, critics charged that the "science" they used to support these positions was intentionally misleading and methodologically flawed. Both works confused correlation with causation (Gould, 1981/ 1996; Heckman, 1995). In addition, *The Bell Curve* (1994) depended upon simple regression analysis to examine complex social issues, and used a narrow definition of SES as the proxy for all environmental experiences (Heckman, 1995; Rury, 1995). Furthermore, the authors of *The Bell Curve* (1994) occasionally omitted the magnitude of correlations used to support their conclusions when the strengths of the correlations were low (Gould, 1981/ 1996; Heckman, 1995). Finally, both works tended to rely heavily on likeminded,

sometimes discredited writers (Kamin, 1999; Lane, 1975/ 1999: Ryan, 1999), were more critical of research that contradicted their positions (Rury, 1995), and misrepresented research that presented opposing viewpoints as supporting their conclusions (Brace & Livingstone, 1971/ 1975/ 1999 on Jensen; Bronfenbrenner, 1972/ 1975/ 1999 on Jensen; Kamin, 1999 on the Bell Curve).

Herrnstein and Murray (1994, pp. 122-125) warned of the danger of inferring causation from correlations, but both Jensen and Herrnstein and Murray treated correlations as though a causal relationship existed (Heckman, 1995). For example, Herrnstein and Murray (1994, Chapter 5) concluded that IQ caused poverty simply because there was a stronger correlation between their measure of IQ and their index of socio-economic status. Not only was it possible that social or economic success contributed to scores on intelligence tests, it was also possible that IQ was merely a contributing factor to this success (Gladwell, 2008; Gordon & Green, 1974/ 1975/ 1999; Rury, 1995), or that some other factor or factors contributed to both IQ and social or workplace success (Kamin, 1999). Indeed, they failed to consider a number of factors unrelated to IQ that might contribute to success. Such factors included, but were not limited to, personality, motivation, interest, sociability, and luck (Gladwell 2008; Gordon & Green, 1974/ 1975/ 1999; Heckman, 1995; Rury, 1995).

This possibility that another factor may contribute to social or workplace success was supported by the fact that neither their index of SES nor their measure of IQ explained much of the variance in measures of success (Gould, 1981/ 1996; Heckman, 1995). For example, Heckman pointed out that Herrnstein and Murray's (1994) measure of IQ never explained more than 30% of the variation in wages.

The measures of environmental effects used by both Jensen (1969) and Herrnstein and Murray (1994, p. 123, Appendix 2) were also problematic. Both used a narrow index of SES as a proxy measure of environmental experiences (Heckman, 1995; Rury, 1995). For example, Herrnstein and Murray (1994) used parental occupation, education and income *at a single point in time* to summarize 15-23 years of experiences (Heckman, 1995). Indeed, for many of the subjects included in Herrnstein and Murray's measures, parental income was not reported, and therefore was omitted from the equation. Furthermore, while SES could provide clues as to the likelihood that a child was exposed to books in the home, it did not provide measure of the degree of parental warmth, exposure to racism, or the many other subtle factors that may have affected IQ scores (Biesheuvel, 1972/ 1975/ 1999).

Indeed, both Heckman (1995) and Rury (1995) noted that in narrowing their measure of environment to a simple index of SES, Herrnstein and Murray (1994) all but ignored the effects of education on cognitive ability. This was no small oversight, since IQ tests were designed to predict educational attainment, because there was such a high correlation between IQ and educational attainment, because a number of studies have shown increased mean IQ scores with improvements in educational systems (Montagu, 1945; Rury, 1988, 1995), and because IQ scores have been shown to increase with additional years of schooling (Heckman, 1995). For example, Neal and Johnson (1994) utilized the same measure of intelligence as that used by Herrnstein and Murray (1994) to demonstrate that an additional year of schooling accounted for a .22-.25 standard deviation rise in IQ scores.

Critics charge that both "How Much Can We Boost IQ and Scholastic Achievement" (1969) and *The Bell Curve* (1994) had other serious methodological flaws. For example, Heckman (1995) noted that Herrnstein and Murray (1994) were inconsistent in whether they

used R^2 or statistical significance as their preferred method of demonstrating causation. While they claimed that statistical significance was the better measure, had they used statistical significance consistently they would have had to admit to a factor in addition to *g* to explain scores on their intelligence test.

In addition to problems in their chosen statistical procedures, critics charged that the authors of both works relied heavily on likeminded, and at times discredited writers to support their arguments (Lane, 1975/ 1999 on the Bell Curve; Montagu, 1975/ 1999c; Ryan, 1999 on the Bell Curve). Indeed, at least one critic noted that Herrnstein and Murray (1994, Chapter 13) relied heavily on researchers who had received funding from the Pioneer Fund, "a nativist, eugenically oriented" endowment (Kamin, 1999 p. 400). Many of the authors cited in *The Bell Curve* also had connections to *Mankind Quarterly*, a journal founded to demonstrate "the genetic superiority of the white race," (Lane, 1975/ 1999). Indeed, as of 1999, Arthur Jensen had received over 1.1 million dollars from the Pioneer Fund, and was cited in The Bell Curve 23 times (Lane, 1975/ 1999). Richard Lynn, whom Herrnstein and Murray described as "a leading scholar on racial and ethnic differences," (Herrnstein & Murray, 1994, p. 272) sat on the boards of both the Pioneer Fund and *Mankind Quarterly*, and had himself received $325,000 as of 1999 (Lane, 1975/ 1999; Southern Poverty Law Center Web site, "The Pioneer Fund" Retrieved on December 7, 2013 from http://www.splcenter.org/get-informed/intelligence-files/groups/pioneer-fund).

At times, Jensen (1969, pp. 52, 76) and Herrnstein and Murray (1996, Chapter 13) both cited researchers who distorted or misrepresented the findings of other researchers so as to support their arguments (Brace & Livingstone, 1971/ 1975/ 1999 on Jensen; Bronfenbrenner, 1972/ 1975/ 1999 on Jensen; Kamin, 1999 on the Bell Curve). For example, Jensen cited Honzik

(Honzik, 1957) who provided a distorted interpretation of the work of Skodak and Skeels (1949). Skodak and Skeels found that the IQs of adopted children were highly correlated with that of their biological mother, but not correlated with their adopted mother. Furthermore, they found that as the children grew older, their IQ scores became more like their biological mother's. However, Honzik (1957), and subsequently Jensen (1969), omitted Skodak and Skeels's (1949) finding that the adopted children averaged IQ scores that were twenty points higher than that of their biological mothers. Skodak and Skeels (1949) interpreted their findings as providing evidence that IQ scores improved when children were placed in better environments. In addition, Skodak and Skeels (1949) concluded that selective placement by adoption agencies resulted in children whose biological mother's had higher IQs being placed in the more nurturing foster homes. In writing his paper, Jensen chose to ignore the interpretations of the original authors in favor of those of Honzik (Brace & Livingstone, 1971/ 1975/ 1999; Bronfenbrenner, 1972/ 1975/ 1999).

Likewise, Herrnstein and Murray (1994, Chapter 13) cited a study by Richard Lynn (1991) wherein Lynn converted Raven's Matrices scores into "bogus" IQ scores (Kamin, 1999). Kamin reported that Raven repeatedly warned that such conversions were not valid, since the matrices distribution was not a normal distribution (see, for example, Raven, Summers, Birchfield, Brosier, Burciaga, Bykrit, et al., 1990). Furthermore, like Jensen's earlier citation of Honzik, Herrnstein and Murray (1969) repeated distortions and omissions made by Lynn as he reported on the work of other researchers (Kamin, 1999).

Finally, Herrnstein and Murray (1994) had a tendency to be highly critical of the research methodologies of papers that reached conclusions opposing their views, while being

relatively uncritical of research methodologies of papers that supported their viewpoint (Rury, 1995).

The Nature of Intelligence

Jensen (1969, p. 9), and Herrnstein and Murray (1994, p. 22) took pains to suggest that *g* was established beyond doubt, and Herrnstein and Murray seemed to believe that the existence of *g* was critical to their argument. However, both Heckman (1995) and Gould (1981/ 1996) argued that *g* had neither been established, nor was the existence of *g* all that relevant to the debate. *g* was a measure of how scores on varied tests related to each other. Heckman (1995) and Gould (1981/ 1996) both argued that *g* was an artifact of linear correlation analysis. A *g* score could always be constructed that explained relationships among scores across a battery of tests. Furthermore, *g* explained 55-71% of the variation in test scores (Heckman, 1995). While this was certainly significant, it suggested the existence of a secondary factor that contributed to test scores. Indeed, Heckman pointed out that despite Herrnstein and Murray's (1994, pp. 15-19, 22) claim to the contrary, psychometricians were divided over the existence of *g*. Noting this and several theories suggesting that intelligence was a composite of multiple traits, Heckman (1995) concluded that there was not much evidence to support *g*.

Whether a mathematical artifact or a factor that determined intelligence, the real test of *g* was whether it predicted human behavior (Heckman, 1995). Heckman found "much evidence" that *g* predicted educational, social, and job related behavior. However, there was also much evidence that other tests and other variables were required to explain the remaining variation. For example, Mincer (1972) found that education and job experience explained much of the variation in wages not explained by *g*.

Understanding of Modern Genetic Theory

Other critics charged that neither Jensen (1969) nor Herrnstein and Murray (1994) had a good understanding of genetics (Gould, 1974/ 1975/ 1999; Montagu, 1975/ 1999a, 1975/ 1999c). Gould (1981/ 1996, 1999) pointed out that to the biologist, "inherited" did not mean "inevitable." It merely meant that something was passed from parent to child. Certain environmental conditions had to exist for an inherited trait to be manifest. Furthermore, modern genetic theory held that genes always interacted with other genes and with the environment to produce the observed trait, or phenotype (Gould, 1974/ 1975/ 1999; Lewontin, 1970/ 1975/ 1999).

Lewontin, (1970/ 1975/ 1999) stated that "Every character of an organism is the result of a unique interaction between the inherited genetic information and the sequence of environments through which the organism has passed during its development," (Lewontin, 1970/ 1975/ 1999, p. 239). Indeed, he noted that each genotype had its own IQ (phenotype) distribution based upon the unique sequence of environmental experiences that each individual passed through.

Finally, some critics (Gould, 1981/ 1996; Sanday, 1972/ 1975/ 1999) suggested that intelligence was probably polymorphic, meaning that multiple genes contributed to intelligence, and each gene was likely affected by the environment in different ways and varying degrees. Therefore, inherited intelligence was the outward manifestation of a wide array of chance distributions and complex interactions.

Gould (1974/ 1975/ 1999) summarized his position on genetically inherited intelligence by stating, "I do not claim that intelligence, however defined, has no genetic basis—I regard it as

trivially true, uninteresting, and unimportant that it does. The expression of any trait represents a complex interaction of heredity and environment. Our job is simply to provide the best environmental situation for the realization of valued potential in all individuals," (Gould, 1974/ 1975/ 1999, p. 188).

To be sure, a number of critics pointed out that it was inappropriate for Jensen (1969, for example see pp. 51-59; 78-88) and Herrnstein and Murray (1994, pp. 105-108, Chapter 13) to use a heritability coefficient calculated on one population to make inferences about another population (Brace & Livingstone, 1971/ 1975/ 1999; Dobzhansky 1973; Gould, 1974/ 1975/ 1999, 1981/ 1996; Lewontin, 1970/ 1975/ 1999). As stated earlier, Jensen's (1969) heritability coefficient was calculated from studies of monozygotic twins. Because the genotype of monozygotic twins did not vary, heritability could be calculated by determining the portion of the variation in IQ scores due to different environmental exposures, and subtracting this ratio from 1.00. A trait with a heritability coefficient close to 0.00 was not considered to be heritable, while a trait with a coefficient close to 1.00 was considered to be highly heritable. Based on several studies of monozygotic twins reared apart, Jensen (1969, pp. 51-59) estimated that the heritability coefficient for intelligence was 0.80.

The problem with this reasoning was that the heritability coefficient was protean: it varied depending on the differences in the ranges of environmental conditions to which members of the respective populations were exposed (Biesheuvel, 1972/ 1975/ 1999; Bronfenbrenner, 1975/ 1999; Heckman, 1995; Lewontin, 1970/ 1975/ 1999). Gregg and Sanday, (1971) noted that heritability was not a measure of the magnitude of the genetic contribution of a trait. Rather it was the "extent to which variability in a trait is due to genetic factors relative to environmental factors," (Sanday, 1972/ 1975/ 1999, p. 280). A homogeneous environment resulted in a

heritability coefficient closer to 1.0, since it was more likely that variations in IQ scores were caused by genetic variation rather than variations in the environment. Conversely, a heterogeneous environment would result in a heritability coefficient closer to 0.0, since more of the variation would be due to different environmental exposures. Therefore, a heritability coefficient was specific to a particular population that experienced a specific range and sequence of environmental conditions.

Because monozygotic twins had identical genotypes, any differences in IQ scores had to be due to the environment. Monozygotic twins reared apart experienced different environments, allowing for heritability to be estimated. However, critics suggested that most of the twin studies that Jensen used for his heritability estimate involved white, working class children from Great Britain (Bodmer, 1972/ 1975/ 1999; Gordon & Green, 1974/ 1975/ 1999). These children came from a narrow range of environmental conditions. Furthermore, most adoption agencies, including those involved in these studies, tended to serve a relatively homogeneous clientele, resulting in selective placements of monozygotic twins into adopted homes that varied little (Bodmer, 1972/ 1975/ 1999; Bronfenbrenner, 1972/ 1975/ 1999). Jensen's error was in applying these heritability coefficients to children who experienced a much broader range of environmental conditions than those on which the heritability coefficients he used had been calculated.

Furthermore, Lewontin (1970/ 1975/ 1999) noted that sometimes very subtle differences in environment had large impacts on phenotype. For example, the absence of a critical micronutrient could have a dramatic impact on cognitive ability. Similarly, people were treated differently based upon superficial differences in appearance, including skin color and hair texture. Such treatment could have significant effects on the mean IQ score of that population.

The social and physical environment to which individuals and populations were exposed could vary in many complex and often subtle ways, and it was nearly impossible to control for all possible environmental variations in human populations (Bodmer, 1972/ 1975/ 1999; Rury, 1995).

Race

A number of critics charged that Herrnstein and Murray (1994, pp. 271) also demonstrated their ignorance of genetics in their treatment of "race." Herrnstein and Murray (1994) admitted that "race" was a socially defined construct, and used the self-identification of subjects to study "race." To be sure, most studies of "race" and intelligence used superficial, observable traits to identify subjects by "race," such as skin color or cultural characteristics (Fried, 1968). They were not based upon genetic patterns that differentiated "races" (Lieberman, Littlefield & Reynolds, 1999). Furthermore, as Montagu (1975/ 1999a, 1975/ 1999b) pointed out, it was unreasonable to assume that the genes that helped determine such superficial characteristics as skin color or hair texture also determined intelligence.

Indeed, there was a debate over whether "race" was even a valid construct (Lieberman, Littlefield & Reynolds, 1999). While a number of writers argued that there were genetic patterns that distinguished one "race" from another (See, for example Garn 1964 and Coon 1962), Montagu (1941) noted that "race" was not an inherited complex of genes that differentiated one "race" from another by appearance, culture and intelligence. Rather, some genes appeared more frequently in specific populations, but the genes associated with any "race" could be found in all populations (Bodmer, 1972/ 1975/ 1999; Garn, 1964).

These critics argued that, in terms of genetics, it was more appropriate to study "races" in terms of clines and frequency distributions. Clines referred to the genetic gradations that occurred between adjacent populations due to inbreeding (Bodmer, 1972/ 1975/ 1999; Huxley, 1938). As such, there was no distinct genetic or geographical boundary between "races," nor was there a distinct genetic definition for any specific "race." Distribution frequencies were the frequencies with which certain genes and gene combinations occurred in different populations (Bodmer, 1972/ 1975/ 1999; Lewontin, 1970/ 1975/ 1999; Montagu, 1975/ 1999c).

In addition, Brace (1964) noted that natural selection operated in ecological zones that did not necessarily coincide with racial boundaries. Studies of the geographic distribution frequencies of genetic markers associated with specific "races" showed that they were not always distributed in the same geographic directions (Bodmer, 1972/ 1975/ 1999). For example, genetic markers related to skin pigmentation tended to vary from North to South, whereas the frequency of "race" related blood markers such as sickle cell trait tended to vary from East to West.

Finally, the classification of people by "race" depended upon the whim or personal history of the classifier more than any pattern of gene frequencies. Therefore, it was more appropriate to shift the discussion of "race" to cultural definitions of ethnic groups (Boas, 1940).

Validity of Tests

Not only did critics take issue with the genetic basis for the conclusions of Jensen (1969) and Herrnstein and Murray (1994), they also questioned whether the tests actually even measured intelligence (Gould, 1974/ 1975/ 1999; Montagu, 1975/ 1999a). Biesheuvel (1972/ 1975/ 1999) noted there was not yet a common definition of intelligence that was defined in

terms of human behavior. Gould (1974/ 1975/ 1999) argued that intelligence tests measure something that equated to success in school, but questioned whether that something might include traits other than intelligence. Montagu (1975/ 1999c) argued that the cumulative research over the preceding 50 years suggested that intelligence might best be defined as "a large assembly of highly varied, overlapping adaptive abilities or skills, rather than…a single faculty; that is, indeed, largely the summation of the learning experiences of the individual," (Montagu, 1975/ 1999c, p. 190). He also noted that the tests always measured learning, and it was therefore impossible to parse out original capacity. Even Jensen noted that intelligence was defined by the test (Jensen, 1969, p. 8), as well as the correlation of the tested abilities to scholastic and occupational success in industrialized societies (Jensen, 1969, p. 14), suggesting that IQ was culture bound. However, Jensen was arguing that certain "races" were genetically predisposed to succeed in a culture that valued these specific inherent traits.

While not specifically questioning whether the Armed Forces Qualification Test (AFQT) actually measured intelligence (Herrnstein & Murray, 1994, pp. 73-74), Heckman (1995) stated that the AFQT was not a measure of *g*, as Herrnstein and Murray defined *g*. Furthermore, as discussed earlier, Neal and Johnson (1994) demonstrated that an extra year of schooling raised AFQT scores. This suggested that some of the factors measured by the AFQT were subject to environmental manipulation, or learning.

Similarly, some critics suggested that IQ tests measured cultural traits rather than innate abilities. Sanday (1972/ 1975/ 1999) noted that geographically isolated groups tended to perform worse on IQ tests than did the mainstream population. After noting that hereditarians interpreted this to be due to genetic isolation, she suggested that the isolation of these groups prevented their exposure to "cultural elements related to the expression of IQ," (Sanday, 1972/ 1975/ 1999, p.

294). The advantage of this hypothesis was that it explained phenomena such as the lower IQ scores of deaf children or children raised in orphanages. Such phenomena were not explained by the hereditarian hypothesis.

Additionally, Davis (1948) defined "problem solving" as a series of learned acts. Problem solving strategies measured by the tests were likely skills that the subject had learned.

Several writers suggested that the tests were not only culture bound, but biased in favor of white middle class culture (Biesheuvel, 1972/ 1975/ 1999; Kagan, 1975/ 1999; Sanday, 1972/ 1975/ 1999). Kagan (1975/ 1999) pointed out that intelligence tests were developed by middle class white men, and Montagu (1975/ 1999c) noted that the tests were arbitrarily standardized on middle class white children. Eells, Davis, Havighurst, Herrick, and Tyler, R. W. (1948) conducted an item analysis of tasks taken from IQ tests and found that they involved content to which only middle and upper class children were likely to be exposed. Finally, Kagan (1975/ 1999) described five classes of questions that appeared on intelligence tests. These included vocabulary, problem solving, analogies, picture completion, and memorization of number sequences. He described how all but memorization of number sequences favored individuals with extensive knowledge of middle class white culture.

Compensatory Education

At the heart of both "How Much Can We Boost IQ and Scholastic Achievement" (Jensen, 1969, pp. 2, 96-109) and *The Bell Curve* (Herrnstein & Murray, 1994, Chapters 8, 9, 14, 15, 17, and 19) was the position that attempts to meliorate social problems through welfare, compensatory education, or affirmative action were futile, and should therefore be abandoned. However, Heckman (1995) argued convincingly that Herrnstein and Murray's (1994) arguments

over the existence of *g*, the role of intelligence in causing social problems, and the inherent and immutable nature of intelligence were irrelevant to their ultimate claim that social programs were ineffective, at best. This was because they never conducted a cost benefit analysis on the social programs in question. Even if such programs had no effect on intelligence, other factors contributed to social and workplace success. To dismiss these programs out of hand without first assessing the outcomes for individuals relative to the costs of the programs reduced the debate to one over ideology rather than empirically demonstrated effectiveness. Indeed, recent research by Heckman and his colleagues have demonstrated that compensatory programs improve opportunities for children by operating on factors other than intelligence (Heckman and Kautz, 2013; Heckman, Pinto, and Salveyev, 2013).

Other critics also took issue with Jensen's (1969) and Herrnstein and Murray's (1994) treatment of social programs. Their arguments generally took three forms. First, even if intelligence was primarily inherited, the conclusion that inherited traits were fixed demonstrated a lack of understanding of genetics (Gage 1972; Gould, 1974/ 1975/ 1999; Lewontin, 1970/ 1975/ 1999). Second, both "How Much Can We Boost IQ and Scholastic Achievement" (Jensen, 1969, pp. 96-109) and *The Bell Curve* (Herrnstein & Murray, 1994, Chapter 17 and 19) misrepresented the research on compensatory programs. Such programs actually produced changes in IQ and achievement, although more research was needed to maximize such gains (Barnett, 1995; Bronfenbrenner, 1974/ 1975/ 1999; Lee, Brooks-Gunn, Schnur, Liaw, 1990; Ryan, 1999). Finally, even if compensatory programs as they then existed failed to have strong or lasting effects, society had a moral obligation to continue to search for and implement interventions that would improve the chances for academic, social, and workplace success

among poor and minority children (Bodmer, 1972/ 1975/ 1999; Bronfenbrenner, 1972/ 1975/ 1999; Gould, 1974/ 1975/ 1999; Lewontin, 1970/ 1975/ 1999; Luria, 1974/ 1975/ 1999).

Several critics noted that even if intelligence were entirely genetic, this did not mean that it was immutable (Gage, 1972; Gould, 1974/ 1975/ 1999; Lewontin, 1970/ 1975/ 1999). There were many examples in which environmental modifications meliorated problems caused by genetically determined conditions. Individuals with myopia could wear prescription glasses. Diabetics could take insulin. At one time, there were no treatments for such conditions. Similarly, even if current compensatory programs had small or transient effects on the IQ of children, this only meant that more research was needed to find effective and permanent interventions.

Some critics also charged that both Jensen (1969, pp. 96-109), as well as Herrnstein and Murray (1994, Chapter 17) misrepresented research when they suggested that compensatory programs had only negligible, transient effects, and that any gains from such programs did not justify the costs. For example Ryan (1999) noted that while Head Start had not lived up to expectations, it was less a failure than Herrnstein and Murray made out. Furthermore, these transient effects of compensatory programs were merely evidence that there was no inexpensive easy fix for early environmental deprivation.

Bronfenbrenner (1974/ 1975/ 1999) examined a variety of compensatory programs from the late 1960s and early 1970s. He concluded that compensatory programs could have significant effects on IQ, such effects tended to fade when the program ended, the fading of these effects was more dramatic for the most disadvantaged children, and programs that focused on building capacity of the child's primary caregivers had stronger and more lasting effects.

Lee, Brooks-Gunn, Schnur, and Liaw (1990) looked just at Head-Start Programs and their long-term impact on disadvantaged Black children. They found that the programs had significant impacts on the IQ scores of these children. While these effects diminished over time, children participating in these programs maintained advantages in cognitive functioning over children who did not participate in such programs. The authors attributed Head Start's diminishing effects on cognitive function to the fact that the children's' home environments remained deficient.

Barnett (1995) reviewed 36 early childhood intervention programs and found that such programs could have significant short term effects on IQ, and significant long-term effects on student achievement. However, he noted that such effects varied with the quality of the program.

In a comprehensive review of the research on the effects of early intervention on cognitive development, Phillips and Shonkoff (Eds.) (2000) concluded that the explosion of research in the latter two decades of the twentieth century demonstrated "(1) the importance of early life experiences, as well as the inseparable and highly interactive influences of genetics and environment, on the development of the brain and unfolding of human behavior; (2) the central role of early relationships as a source of either support and adaptation or risk and dysfunction; (3) the powerful capabilities, complex emotions, and essential social skills that develop during the earliest years of life; and (4) the capacity to increase the odds of favorable developmental outcomes through planned interventions," (pp. 1-2). They noted that the debate over nature versus nurture was obsolete, and that mainstream scientists took the position that genes interacted with the environment to determine children's cognitive abilities.

Finally, several critics argued that even though finding effective interventions was challenging, society had a moral obligation to continue trying. Lewontin (1970/ 1975/ 1999)

noted that even if no intervention was shown to increase IQ at present, this did not mean that no such intervention was possible. This would be akin to early 20th century scientists abandoning attempts at human flight because humans had never before flown. Finally, several critics (Bodmer, 1972/ 1975/ 1999; Bronfenbrenner, 1972/ 1975/ 1999; Gould, 1974/ 1975/ 1999; Luria, 1974/ 1975/ 1999) noted that society had an obligation to provide the environmental conditions necessary to help each child to realize his or her full genetic potential.

Conclusion

In 1924, a young Horace Mann Bond (1924a) exhorted African American students to educate themselves so that they would be prepared to defend against those who would disguise racism in a cloak of science. In 1956, Bond responded to a coalition of politicians who used "scientific data" from the World War I Army Testing Program to fight school desegregation. In this instance, Bond used the same "scientific" arguments to show that Southern politicians were intellectually inferior to those from the North. His point was not to prove that Southern politicians really were "feebleminded," but to demonstrate how the "science" used to support the racially based caste system was a distortion of impartial scientific methods. In his classic work *The Mismeasure of Man*, Gould (1981/ 1996) provides a quote by Alfred Binet that is still relevant. Binet wrote, "It is really too easy to discover signs of backwardness in an individual when one is forewarned," (Binet, 1905, as cited in Gould 1981/ 1996). It is also far too easy to take statistics at face value when they support *a priori* held beliefs.

In his response to *The Bell Curve*, (Heckman, 1995) noted that much of the evidence cited in the document was empirically correct. Psychometrics was a legitimate science. Tests were not culturally biased. Different demographic groups varied according to their mean scores

on intelligence tests. Test scores were relatively stable by the time the individual reached young adulthood. And, a linear combination of test scores such as would produce a *g* did predict work place productivity and social performance. However, this did not prove the existence of *g*, nor was *g* central to their argument. Different combinations of test scores were also predictive of academic, workplace, and social performance. In addition, variables other than test scores were necessary to explain variations in academic, workplace, and social performance. Furthermore, demonstrating differences in ability at age 17 was a far cry from proving that those differences were primarily due to differences in genetic material rather than differences in experiences. Finally, regardless of the ratio of the effect of the genetic relative to the environmental components on IQ scores, there was an environmental effect. Without doing a cost/ benefit analysis of compensatory programs or affirmative action on academic, workplace and social performance, the book was pointless.

Following the publication of *The Bell Curve*, Linda Gottfredson wrote an editorial to the *Wall Street Journal* entitled *Mainstream Science on Intelligence* in which she listed much of the empirical evidence surrounding intelligence and intelligence tests (Gottfredson, December 13, 1994). However, this article was a show of support for the interpretations of evidence evidence provided by Herrnstein and Murray. The document was signed by 52 "mainstream scientists," including Richard Lynn, Robert Gordon, Hans Eysenck, J. Phillippe Rushton, and Arthur Jensen.

Interpreting the empirical evidence in ways that did not flow from the evidence not only supported "race" and class based hegemony in the United States, it had a more insidious effect. The use of science and statistics to provide evidence in support of a "race" and class based caste system communicated a message of inferiority that could be internalized by those at the bottom of the racial hierarchy. In discussing the effects of segregated schools on African American

children, the Warren court ruled that to separate African American children "from others of similar age and qualifications solely because of their race generates a feeling of inferiority as to their status in the community that may affect their hearts and minds in a way unlikely ever to be undone," (*Brown v. Board of Educ*., 1954, p. 494). The use of correlations to infer causation, group means, heritability formulas and narrow definitions of environment also communicate this message of inferiority. Only this evidence is more direct and more dangerous because it has the imprimatur of science. Bond's warning to arm oneself against those who would enlist science in support of racism is still relevant.

Chapter 5: Standards, Testing and Accountability

The standards, testing and accountability reforms were promoted as a means to improve the quality of education for *all* American children, while closing achievement gaps between traditionally underserved children and their peers. These reforms acted through the establishment of common rigorous standards, and using aligned tests to hold educators accountable for teaching these standards (Finn, Kanstoroom, Rothstein, & Honig, 2001; Smith & O'Day, 1991). This would produce systemic reform by motivating educators to align all aspects of education with the standards (Smith & O'Day, 1991). The debates that were sparked by these reforms will be discussed below, and will cover the following topics: political orientation; critique of public education; business and economic oriented reforms; narrowing of educational outcomes and experiences; equity and equality; dehumanizing language and practices; and meritocracy.

Political Orientation

While not exclusively associated with conservatives, critics characterized these reforms as originating from neoliberal and neoconservative political and philosophical tendencies (Apple, 2001). Neoliberal philosophy elevated individualism over individuals and human rights. Personal freedom was defined as self-determination, the right to hold property and the right to make consumer choices. This philosophy involved a kind of "freedom with responsibility," in the belief that market style competition would hold individuals and businesses accountable for their actions. Through their choices, consumers were better positioned than government to hold individuals and businesses accountable. Therefore, the role of the government was limited to opening new markets, keeping existing markets open, and eliminating regulations that might impede business (Apple, 2001; Harvey, 2005).

A pure neoliberal reform would have involved a network of private schools that competed with each other for customers. Although standards and accountability reforms were not this pure manifestation of neoliberal ideas, these were influential on standards and accountability reform proposals. These reform proposals included neoliberal ideas such as the use of choice as a mechanism for holding schools accountable, the focus on serving the economic and business needs of the country, and the undermining of the power of teachers and teachers' unions (Harvey, 2005).

Similarly, neoconservative philosophy included many of these same beliefs regarding individualism and the power of the markets to hold individuals accountable (Apple, 2001; Harvey, 2005). Neoconservative philosophy also held that regulations impeded business, and therefore infringed on the freedom to hold personal property and make consumer choices. However, neoconservatives recognized that aspects of neoliberal philosophy undermined social cohesion. They addressed this weakness by promoting a strong state that would keep order and by advocating for a return to common "traditional American values."

Therefore, neoconservatives favored the addition of external rewards and punishments as mechanisms for motivating educators to improve. Furthermore, neoconservatives also supported mandated centralized testing, common standards that promoted "traditional American values," and a centralized mechanism for holding schools accountable for student learning (Apple, 2001).

Conversely, proponents of standards testing and accountability sometimes characterized the opposition as coming from the political left, while acknowledging some opposition from conservatives (Finn, Kanstoroom, Rothstein, & Honig, 2001). Finn et al., described opponents as coming from the "Romantic Thoughtworld" which "is nearly always found on the political left," (Finn, et al., 2001, p. 153). However, they acknowledge opposition from conservatives, such as

"libertarians and conservatives who believe ardently in local control, marketplace mechanisms, and school level pluralism…who do not trust the state to decide what children should learn," (p. 153).

Similarly, even as critics characterized standards, testing and accountability reforms as originating with conservative philosophy, many on the left embraced these reforms as a mechanism for improving education for traditionally underserved children (Archer, 2006; Loveless, 2007; Reid, 2005; Salzman, 2006). Loveless wrote that "NCLB consists of conservative ideas—testing, accountability and incentives—wrapped in liberal clothing—a big federal program that seeks, as its primary objective, the equalization not only of educational opportunity but also of educational outcomes," (Loveless, 2007, p. 272).

NCLB (2002) passed with bipartisan support. However, public support for NCLB (2002) was more complex. Loveless (2007) analyzed public opinion surveys conducted both just prior to passage and three years into implementation of NCLB (2002). He found that while Republicans were more likely to support NCLB (2002) and Democrats to oppose it, African Americans and Hispanics strongly supported the law. He concluded that people who felt that schools were failing to serve their children were more supportive of strong education reforms like NCLB (2002).

Historically, poor and minority students in the United States had not been given equal access to the academic preparation required for them to gain admittance to higher education and higher paying occupations. Common rigorous standards would provide poor and minority students the opportunity to learn the same standards as their more affluent peers (Ravitch, 1996). In addition, testing and accountability would hold schools accountable for ensuring that these students achieved minimum learning outcomes compared to these standards (Gamoran, 2007b;

Kearns, 1988; Rury, 2012). Requirements that test scores be disaggregated and published by subgroups would draw attention to inequalities in educational outcomes for these students that had previously been obscured (Gamoran, 2007b; Rury, 2012). Finally, accountability mechanisms that included supplemental education services and opportunities to transfer out of schools needing improvement provided additional educational opportunities to children who attended failing schools (DeBray-Pellot & McGuinn, 2009; Gamoran, 2007b).

Critique of Public Education

Calls for standards, testing, and accountability gained momentum following the publication of *A Nation at Risk: The Imperative for Educational Reform* (Commission for Excellence in Education, 1983; Evers, 2001; Finn, et al., 2001; Gamoran, 2007b). *A Nation at Risk* (Commission for Excellence in Education, 1983) was published in the context of slowing growth in American national productivity and growing trade deficits with America's European and Asian competitors (Rothstein, Jacobson & Wilder, 2008; Sacks, 1999). In addition, rising inflation and stagnant wages meant that more Americans were struggling to get by. Slowed growth, trade deficits, and "stagflation" caused many Americans to question the Keynesian economic policies that had been credited with so much economic growth during the 1940s, 1950s, and 1960s (Harvey, 2005). It was in this context that Ronald Reagan was elected president. Using America's economic troubles as his justification, President Reagan launched a series of neoliberal policy reforms and political appointments aimed at reducing regulations that impeded businesses and international trade, lowering taxes on the wealthiest Americans in order to stimulate economic growth, and gutting social programs (Harvey, 2005).

The authors of *A Nation at Risk* (Commission for Excellence in Education, 1983) argued that America's economic troubles were caused, at least in part, by the decline in the quality of public education. To support this claim, they cited the poor rankings of American students on tests used for international comparisons compared to students from America's economic competitors. The authors also cited the declining average scores on standardized tests such as the Scholastic Aptitude Test (SAT) and the National Assessment of Educational Progress (NAEP) (Commission for Excellence in Education, 1983; Madaus, Russell & Higgins, 2009; Rothstein, Jacobson & Wilder, 2008). Finally, they argued that in order to regain its economic competitive edge, the United States had to improve public education. This improvement could best be achieved by raising academic standards, increasing testing, and holding students and educators accountable for academic achievement.

A Nation at Risk ushered in a wave of attacks on public schooling in America. Critics of American education claimed that American schools were overfunded and failing (Hanushek, 1989, 1994; Kearns, 1988; Walberg, 1994, 1998). As in *A Nation at Risk* (Commission for Excellence in Education, 1983), critics cited the poor rankings of American students compared to students from other countries on tests of international comparisons (Walberg, 1998), and declining NAEP and SAT scores (Hanushek, 1989, 2003; Ravitch, 1999). These criticisms of American education were usually accompanied by calls for improved standards, testing, accountability, and market style reforms (Kearns & Doyle, 1989; Sacks, 1999).

Defenders of public education argued that while there were very real problems with many schools in the United States, the evidence cited by public school critics, in general, and the authors of *A Nation at Risk* in particular, was misleading. Berliner and Biddle (1995) claimed that variations in curriculum caused American school children to have less exposure to the tested

content, giving them a disadvantage on tests of international comparison. Furthermore, they suggested that it was inappropriate to compare decentralized, heterogeneous American schools with centralized homogeneous foreign schools. Other defenders of public education claimed that test scores were an insufficient measure of school quality (Koretz, 2008).

Other indicators, such as the percentage of adults that graduated from college (Sacks, 1999) or the percentage of Americans that had at least some high school (UNESCO 1997) provided evidence that American schools were doing quite well compared to other countries. However, these arguments ignored or downplayed the fact that American public education was, in general, failing poor and minority children.

Public school defenders also argued that tests such as the SAT or ACT were not validated as measures of school quality (Airasian, 1987). Rather, these tests were designed and validated to predict freshman year grades in college (Madaus, Russell & Higgins, 2009).

Furthermore, the SAT and ACT tests were elective. Students were not randomly assigned to take these tests. Defenders of public schools noted that an increased proportion of students from traditionally lower scoring demographic groups began taking these tests in the years leading up to *A Nation at Risk* (Commission for Excellence in Education, 1983). As such, much of the decline in SAT and ACT test scores could be attributed compositional changes in the population taking the test (Berliner & Biddle, 1995; Hanushek, 2003; Koretz, 2008; Madaus, Russell & Higgins, 2009). As the proportion of low scoring subgroups increased, their scores pulled down the aggregate averages, even though the scores disaggregated by subgroups may have remained the same or improved (Koretz, 2008).

Indeed, some defenders of public education declared that much of the criticism was aimed at undermining public education rather than reforming it (Berliner & Biddle, 1995; Kohn,

2004b; Madaus, Russell & Higgins, 2009; Ohanian, 1999; Sacks, 1999). According to this argument, politicians motivated by "contempt for public institutions" and a desire to privatize education, promote school vouchers, and lower taxes could use testing and accountability to discredit or undermine schools (Karp, 2004; Kohn, 2004b; Wood, 2004).

Finally, defenders of public education suggested that criticizing schools allowed policy makers to deflect attention from broader social problems that would otherwise be difficult and expensive to effectively address (Madaus, Russell & Higgins, 2009; Rothstein, Jacobson & Wilder, 2008). In addition, criticizing public schools diverted public attention from political and business activities that might be considered self-serving or unethical (Ohanian, 1999). Teachers, schools, and children were politically weak yet resilient, making them convenient scapegoats (Madaus, Russell & Higgins, 2009; Ohanian, 1999).

The critics of public education countered that the "education establishment" had a vested interest in maintaining the status quo (Evers, 2001; Finn, 2002; Walberg, 1998). The implication was that educators and their unions stood to lose money, power, and possibly their jobs if the effectiveness of educators were open to examination.

Economic and Business Interests

Critics of public education tended toward an economic world view. They emphasized the role of schools in serving the business and economic interests of the United States (Commission for Excellence in Education, 1983; Finn, et al., 2001; Hanushek, 1989, 2005; Kearns, 1988; Kearns & Doyle, 1989). Schools were both the cause and cure of America's economic troubles. Indeed, a fundamental purpose of public education was to prepare a productive workforce with the skills needed by business so that America would regain its competitive dominance in the

world economy. Critics of public education complained that schools were failing to provide students with basic reading and computational skills, forcing businesses to provide workers with remedial education (Kearns and Doyle, 1989).

The defenders of American education countered that there was little relationship between the economy and the quality of the workers produced by schools (Bracey, 1996; Ohanian, 1999; Rothstein, Jacobson & Wilder, 2008). Rather, wealth was created by tax, trade, regulatory, and monetary policies (Rothstein, Jacobson & Wilder, 2008).

Furthermore, several critics complained that public schools should not be treated as though they were "handmaidens to American business," (Sacks, 1999), whose mission was to train workers (Bracey, 1996; Ohanian, 1999). Indeed, preparation for a life of often dull or dangerous work should not be an objective of public education (Bracey, 1996). Rather, schools should prepare students for a rich and thoughtful life (Bracey, 1996), to be good citizens (Bracey, 1996; Ohanian, 2003), to be good people, and to find joy in learning (Ohanian, 1999).

Berliner and Biddle (1995) claimed that the emphasis reformers placed on preparing students for high tech jobs was disingenuous because the fastest rate of job growth was in the service sector. Ohanian (1999, 2000, and 2003) went further, suggesting that higher standards and testing were part of a cynical effort to create gluts of both highly skilled and unskilled labor, thereby driving down wages.

Not only did proponents of testing for accountability make the case that schools should serve America's economic and business interests, testing for accountability reforms were based on a business model (Darling-Hammond, 1994; Finn, 2002; Hanushek, 1989; Kearns, 1988; Kearns & Doyle, 1989; Madaus, Russell & Higgins, 2009; McNeil, 1988a, 1988b; Ohanian, 1999, 2000; Sacks, 1999). These reforms included a hierarchical structure (Darling-Hammond,

1994; Finn, Kanstoroom, Rothstein, & Honig, 2001; Smith & O'Day, 1991); market style reforms and competition (Finn, 2002; Finn, et al., 2001; Kearns, 1988; Kearns & Doyle, 1989; Madaus, Russell & Higgins, 2009; Ravitch, 1994; Sacks, 1999; Walberg, 1994, 1998); measures of profitability and productivity (Finn, 2002; Kearns, 1988; Madaus, Russell & Higgins, 2009; Gerstner, 1994); standardization and quality assurance (Kearns, 1988; Ravitch, 1993; Walberg, 1998); and the allocation of rewards and sanctions based on productivity and product quality (Hanushek, 1989; Kearns, 1988).

In something of a paradox, proponents of standards, testing and accountability reforms empowered teachers by blaming them for the decline of public education (Finn, 2002; Walberg, 1988) and advocating that teachers be held accountable for student achievement (Finn, 2002; Finn, et al., 2001; Hanushek, 1994, 2003; Hanushek & Raymond, 2005; Walberg, 1998). If teachers were responsible for poor student achievement, then it followed that they should be able to affect student learning.

Despite this belief in the ability of teachers to impact student learning, advocates of testing for accountability also argued for stripping teachers of important decision making roles (Madaus, Russell & Higgins, 2009; Ohanian, 1999;). They suggested that teachers were unable or unwilling to make decisions that were in the best interests of students (Finn, 2002; Finn, et al., 2001; Hanushek, 2005; Ravitch, 1999; Walberg, 1998). Indeed, several advocates of standards, testing and accountability claimed that only external rewards and punishments would motivate teachers to work harder to raise student achievement (Finn, 2002; Finn, et al., 2001; Hanushek, 1989, 1994, 2003; Hanushek & Raymond, 2005; Walberg, 1998).

Opponents of testing for accountability suggested that this lack of faith in educators was reflected in a reform effort that was characterized by hierarchical decision making, and the

corresponding deskilling of teachers. The assumption of hierarchical intelligence held that those at the top of the hierarchy made better decisions than those at lower levels (Darling-Hammond, 1994). Standards, tests, and sanctions were all developed and imposed by higher authorities external to schools and classrooms (Finn, et al., 2001; Smith & O'Day, 1991).

Even though teachers were excluded from decisions regarding what content was worthy to teach, proponents of standards, testing, and accountability claimed that teachers would be empowered because they would be allowed to determine *how* they would teach content to students (Finn, 2002; Finn, et al., 2001; Hanushek, 1994, 2005; Ravitch, 1999).

However, their opponents argued that in practice, pressures from the threat of accountability caused many schools and districts to further reduce the control and decision making authority of teachers by providing scripted lessons, mandating they teach commercial test prep materials, requiring drill and practice, or demanding that teachers to teach the test (Darling-Hammond, 1994; Madaus, Russell & Higgins, 2009; McNeil, 2000; Ohanian, 1999; Sacks, 1999). Thus, teachers were deskilled because they were excluded from decisions regarding *what* to teach even as school and district leadership no longer trusted them to make good decisions regarding *how* to teach.

Opponents claimed that teachers were deskilled in other ways, as well. Test scores displaced and devalued teacher judgments about student learning (Kohn, 2004a; Madaus, Russell & Higgins, 2009; McNeil, 2000). In addition, time spent preparing for and taking tests limited the available time for teachers to take advantage of teachable moments, follow their intuition, follow student interests, or attend to the emotional needs of individual students (Karp, 2004; Ohanian, 1999). According to this argument, teachers were, in effect, treated like workers on an assembly line rather than semi-autonomous professionals capable of diagnosing educational

needs and prescribing solutions. According to this perspective, standards and tests emphasized the science of teaching while squeezing out the art (Ohanian, 1999).

In contrast to the paradox of empowerment through accountability, opponents of standards, testing and accountability tended to argue for empowering teachers to make important decisions that had an impact on student learning even as they claimed that many of the factors affecting student achievement were beyond the teachers' control (Darling-Hammond, 2000; Karp, 2004; Kohn, 2000; Koretz, 2008; Madaus, Russell & Higgins, 2009; Rothstein, 2004, 2008; Ohanian, 1999, 2003; Sacks, 1999). Of particular significance was the high correlation between SES and cognitive development (Karp, 2004; Koretz, 2008; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999), and the correlation between SES and test scores (Darling-Hammond, 2000; Karp, 2004; Kohn, 2000; Koretz, 2008; Madaus, Russell & Higgins, 2009; Ohanian, 1999, 2003; Rothstein, 2004; Sacks, 1999;). Indeed, Koretz (2008) suggested that test scores were more symptomatic of factors outside of the school's control than they were the results of the skill and efforts of teachers.

As such, opponents argued, it was disingenuous for lawmakers to hold schools accountable for raising test scores and closing achievement gaps without taking actions to meliorate broader social problems that contributed to them (Karp, 2004; Kohn, 2000; Meier 2004; Ohanian, 1999; Rothstein, 2004, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). These social problems, to name but a few, included poverty, malnourishment, inadequate health care, inadequate housing, and unsafe neighborhoods (Rothstein, 2004, 2008; Wood, 2004). According to the critics of test based accountability, only by crafting policies that addressed the root causes of low test scores could real gains in student achievement be realized (Koretz, 2008).

Indeed, some opponents of testing for accountability suggested that this reform perpetuated existing social inequalities by allowing middle and upper class individuals to rationalize the existing social structure. They could argue that they provided high standards, aligned tests and accountability. The responsibility for poor performance would then be placed on the students and their teachers (Kohn, 2000; Ohanian, 1999; Sacks, 1999). In addition, this could be accomplished without a substantial commitment of resources or more equitable distribution of these resources (Darling-Hammond, 2004; Karp, 2004; Madaus, Russell & Higgins, 2009; Meier 2004; Wood, 2004).

It was true that broader social issues needed to be addressed in order to improve the well-being of all children. Certainly, meliorating these social problems might also go a long way towards improving achievement. However, this line of argument on the part of the opponents of standards, testing and accountability to some degree not only abdicated responsibility for student achievement on the part of the very people whose job it was to teach children, it disempowered teachers by suggesting that they had little impact on student achievement. Furthermore, it supported the argument of some promoters of testing for accountability that educators resisted attempts to hold them accountable for student learning (Evers, 2001; Kearns & Doyle, 1989).

Narrowing of Outcomes and Experience

In addition to their claims that it was misguided to hold schools accountable for meliorating larger social problems, opponents of testing for accountability also warned that these reforms distorted the teaching and learning process. This was based on Campbell's Law, which held that when high stakes were attached to attempts to quantifiably measure a complex social process, the social process was corrupted (Campbell, 1975). In part, this was because typically

only a portion of the complex social process could easily be measured. Campbell (1975) warned that when high stakes rewards and sanctions were attached to performance on only a portion of the social process, agents would be motivated to manipulate the process in ways that artificially inflated the scores on that portion that was measured. What follows is a brief description of the mechanism by which critics claimed that high stakes testing altered what was taught and learned, who was taught, and how they were taught.

Promoters of standards, testing and accountability advocated for including as standards only those educational outcomes that could be easily measured. Otherwise, there would be no way to know whether the standards had been achieved (Finn, et al., 2001).

Unfortunately, critics charged, many worthy objectives of public education could not be defined numerically, and therefore would be left off standards documents (Ohanian, 1999; Rothstein, Jacobson & Wilder, 2008). Indeed, in a historical review of attempts to define desired educational outcomes in the United States, Rothstein, Jacobson and Wilder (2008) identified the eight most commonly suggested educational outcomes from the literature. These included 1) basic academic knowledge and skills; 2) critical thinking; 3) appreciation of the arts and literature; 4) preparation for skilled work; 5) social skills and work ethic; 6) citizenship and community responsibility; 7) physical health; and 8) emotional health. Rothstein, Jacobson and Wilder maintained that of these eight, only "basic academic knowledge and skills" and "critical thinking" could be measured by a standardized test, and most tests did a poor job of measuring "critical thinking." The authors concluded that the part of the curriculum that could be easily measured by a paper and pencil test was only a small portion of the desired educational outcomes. Critics of testing for accountability argued that any assessment of education should

focus on the most important educational goals, not just what was easiest to measure (Bracey, 1987; Rothstein, Jacobson & Wilder, 2008).

Critics also warned that the high stakes attached to numerical scores motivated educators to shift focus from pursuing desired educational outcomes to achieving high test scores (Madaus, Russell & Higgins, 2009). Rewards and sanctions worked as intended if they directed institutions toward activities that were rewarded, and away from activities that were not rewarded or were punished. This argument held that strategies that were effective in boosting scores on standardized tests might do little to promote a deep understanding of content, and may even be detrimental to the achievement of other desired educational outcomes (Madaus Russell and Higgins 2009; Rothstein, 2008).

In addition, critics noted that time available for teaching and learning was limited. As time devoted to content likely to be tested increased, time devoted to important but untested activities and content was squeezed out of the school day (Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008). For example, educators were motivated to focus on the tested content of mathematics and language arts at the expense of untested content such as social studies, science or the arts (Kohn, 2000; Madaus, Russell & Higgins, 2009; McNeil, 2000, 2004; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). Furthermore, educators would replace engaging instructional strategies with strategies such as drill, practice and test prep that were believed to raise test scores (Darling-Hammond, 1990; Kohn, 2000; Madaus, Russell & Higgins, 2009; Ohanian, 1999; Sacks, 1999). Finally, schools would decrease time spent on non-academic, yet socially, emotionally, and cognitively enriching activities such as character education, field trips, recess or lunch time (Bracey, 1996; Madaus, Russell & Higgins, 2009; Ohanian, 1999; Wood, 2004). Critics argued that such negative effects would be greatest on

schools serving poor and minority students, since these schools were more likely to feel pressure from accountability sanctions (Darling-Hammond, 2000; Madaus, Russell, & Higgins, 2009; McNeil, 2000; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004).

Critics of testing for accountability warned that narrowing of the curriculum was taken to an extreme when educators "taught to the test" (Bracey, 1987; Darling-Hammond, 1994; Madaus, Russell & Higgins, 2009; McNeil, 1990, 2004; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004). Test questions were samples from the broader content domain (Koretz, 2008). A student's answers to questions from the sample were used to make inferences about his or her learning in the broader content domain. When teaching and learning was limited to the questions sampled for the test, student test scores would rise without a corresponding mastery of the broader content, giving a false impression of what the students have actually learned (Koretz, 2008; Madaus, 1988; Madaus, Russell & Higgins, 2009).

Promoters of standards, testing and accountability countered that there was nothing wrong with teaching to the test, provided it was a good test (Finn, et al., 2001; Smith & O'Day, 1991). If the items that were tested were important enough to be measured, then it was important that students mastered this material. Furthermore, accountability tests influenced classroom practices. Therefore, a well-made test would influence classroom practices in positive ways (Smith & O'Day, 1991). For example, tests could be designed to measure critical thinking skills. This would motivate teachers to teach in ways that helped students to develop these skills. In addition, the problem of teaching to the test could be eliminated simply by selecting different test questions each year.

Finally, opponents of testing for accountability argued that when accountability schemes focused on the percentage of students scoring above a cut score, the magnitude of the scores

were no longer important. This minimized relatively large gains in test scores of students originally far from the cut score while exaggerating relatively small gains made by students who were originally close to the cut score. Thus, a school with large gains could be judged as less effective than a school that had smaller gains, provided the school with the larger gains failed to move as many students above the cut score. Such accountability schemes also gave educators an incentive to focus efforts on students likely to score near the cut score, at the expense of those students likely to score far from the cut score (Koretz, 2008; Madaus, Russell & Higgins, 2009). In this way, high stakes attached to the percentage of students scoring proficient or above created incentives for educators to leave some children behind.

However, proponents of testing for accountability noted that decisions that distorted education were not inevitable. These decisions were made at the local level, and were not endemic to testing (Rury, 2012). In addition, accountability schemes have been devised that measured the magnitude of student growth, thereby incentivizing the maximization of growth of all students (Kanstoroom and Finn, 1999).

Equity and Equality

An important theme in the standards, testing, and accountability reform debates was the relative importance of resources in educating children. For many of those who promoted standards, testing and accountability reforms, public education was overfunded compared to the returns on the investment. Indeed, according to advocates of this reform, there was little evidence that increasing resources improved student learning (Finn, 2002; Hanushek, 1989, 1994, 2003, 2005; Hanushek & Raymond, 2005; Kearns, 1988; Walberg, 1994). Rather, raising standards, testing student learning, and holding schools accountable could provide a means of increasing

outputs in the form of student achievement, with relatively little increase in *inputs* in the form of resources.

For testing advocates, *The Coleman Report* (Coleman, et al., 1966) started a shift from input driven educational reforms, toward output oriented reforms (Finn, 2002; Hanushek, 2003; Madaus, Russell & Higgins, 2009). The authors of *The Coleman Report* concluded that school quality was not as important as student family background in determining student achievement. Because *The Coleman Report* measured school quality in terms of "resources available" and "facilities," testing advocates interpreted its findings as demonstrating that resources did not improve student outcomes (Finn, 2002; Hanushek, 2003).

Opponents of testing for accountability had a different interpretation of *The Coleman Report* (Coleman et al., 1966; Darling-Hammond, 2000; Koretz, 2008). They argued that because there was such a high correlation between student socio-economic background and the resources provided to schools in the United States, it was nearly impossible to separate the effects of school resources on student achievement from the effects of family resources.

However, proponents of testing for accountability cited other evidence to support their contention that additional resources had little effect on achievement (Hanushek, 1989, 1994, 2003, 2005; Hanushek & Raymond, 2005; Kearns, 1988; Walberg, 1994). Increases in spending during the latter half of the 20th Century were not accompanied by equally dramatic increases in student achievement (Hanushek, 1989, 1994; 2003, 2005; Kearns, 1988; Walberg, 1994). Specifically, the increases in spending that fueled rising teacher salaries and decreased class sizes were accompanied by stagnant or declining test scores (Hanushek, 1989, 1994; 2003).

Furthermore, proponents of standards, testing and accountability argued that funding targeted at improving achievement for historically underserved student populations, known as

compensatory education, was ineffective and potentially harmful (Hanushek, 1989; Ravitch, 1999; Walberg, 1998). Walberg (1998) claimed that compensatory programs established a two-tiered education system that labeled and stigmatized students, and denied them access to the core curriculum. These programs failed to close achievement gaps between the targeted students and the rest of the student population, and any improvements in achievement resulting from compensatory programs were short-lived, at best (Ravitch, 1999; Walberg, 1998).

However, several of the opponents of standards, testing and accountability argued that it was disingenuous to hold schools accountable for the performance of poor and minority children when inequitable school funding policies in the United States contributed to the achievement gaps between these students and students who were more affluent or white (Darling-Hammond, 1994, 2000, 2004; Karp, 2004; Kohn, 2000; Ohanian, 1999, 2003; Rothstein, 2004, 2008; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004). The amount of per pupil spending in schools varied widely between states, within states, and even between schools in the same district (Darling-Hammond, 2000, 2004; Rothstein, Jacobson & Wilder, 2008).

According to Darling-Hammond (2004) within state spending differences were as great as three or four to one. Between states, the differences were as high as ten to one. She noted that in 2004 the poorest public schools in the nation received only around $3000 per pupil annually while the wealthiest schools in the nation received around $30,000 per pupil, annually. In addition, schools serving the poorest children tended to be funded at the lowest levels (Kozol, 1991). Even within districts, those schools serving the highest concentrations of minority children often received the fewest resources (Darling-Hammond, 2000). Critics of testing and accountability argued that educational reform should begin with equitable funding for schools (Darling-Hammond, 2004; Karp, 2004; Kohn, 2000; Meier 2004).

According to this reasoning, resources had a direct impact on opportunity to learn. Opportunity to learn relied not just on access to content, it also depended upon access to high quality instruction, adequate and appropriate instructional materials, safe and healthy facilities equipped to support the curriculum, and highly effective teachers (Berliner & Biddle, 1995; Darling-Hammond, 1990, 1994, 2000, 2004; Finn, et al., 2001; Kohn, 2000; Sacks, 1999). Under-resourced schools had less money to spend on teaching staff, resulting in fewer teachers, fewer course offerings, and larger class sizes (Darling-Hammond, 2004). Furthermore, under-resourced school districts had difficulty offering the competitive wages necessary to compete with surrounding districts for high quality teachers (Darling-Hammond, 1994, 2000). Fewer good teachers reduced access to more rigorous content (Darling-Hammond, 2000). Low wages and fewer teachers also increased the likelihood that students would be taught by someone teaching outside of his or her area of certification (Darling Hammond 1994). According to Darling-Hammond (2000, 2004), high quality teachers had the greatest impact on opportunity to learn, but were the least equitably distributed resource.

Similar to their effects on narrowing the educational experiences of poor and minority children, critics claimed that sanctions associated with standards, testing, and accountability reforms disproportionately reduced resources available to the already under-resourced schools serving these children. Sanctions that allowed student transfers, required failing schools to pay for tuition and transportation for these transfers, required failing schools to pay for supplemental education services, or otherwise withheld resources, disproportionately reduced resources available to schools serving the neediest children (Darling-Hammond, 2004; Rothstein, Jacobson & Wilder, 2008; Wood, 2004).

In addition, such sanctions fell harder on teachers and administrators who chose to work with traditionally low scoring populations. This could discourage highly qualified and dedicated professionals from working in high poverty schools, further restricting the opportunities to learn for students attending these schools (Darling-Hammond, 1994, 2004; Kohn, 2000; Ohanian, 1999; Sacks, 1999; Wood, 2004).

Proponents of standards, testing and accountability countered that opportunity to learn would be made more equitable through holding all schools accountable for teaching the same set of common, rigorous standards (Ravitch, 1996). "Standards establish the principle that all students should encounter the same educational opportunities and the same performance expectations, regardless of who their parents are, or what neighborhood they live in (Ravitch, 1996, p. 134; see also Kearns, 1988; Ravitch, 1993; Smith & O'Day, 1991).

Common, academically oriented standards created additional opportunities for poor and minority students (Finn, et al., 2001; Kearns, 1988; Ravitch, 1993). Historically, these students were denied access to content required for entrance into higher education. Mandating that all students have access to these same academically oriented standards would create opportunities for traditionally underserved children to gain entrance to higher education and the high paying professional opportunities that required a college education (Kearns, 1988). As David Kearns stated, "Academic standards for all students have to be raised. How dare we arbitrarily consign our kids to two different futures. [*sic*] Why do the affluent get to go to college prep, while the disadvantaged get dumped into dead-end vocational or general courses," (Kearns, 1988, p. 568).

Equal opportunity to learn was not the only advantage to having common standards. Providing all students with the opportunity to learn rigorous curriculum would improve education for all students in the United States, increasing academic and economic

competitiveness. Therefore, the establishment of rigorous standards was the first step in education reform (Ravitch, 1993). Standards defined what all students should know and be able to do, and the level of performance that should be expected (Ravitch, 1996). Furthermore, standards provided coherence to a fragmented curriculum that varied not only by region and school, but by classroom (Ravitch, 1993; Walberg, 1998). Common standards ensured continuity in education, from one grade to the next, even if the student changed schools (Ravitch, 1993; Walberg, 1998). Standards based reforms were intended to control the content taught in classrooms regardless of the teacher (Ravitch, 1993; Walberg, 1998). Thus, standards established consistency and continuity of what children would learn regardless of the neighborhood in which they lived, the school they attended, or the teacher who taught them.

Proponents of standards, testing and accountability believed that holding educators accountable for students learning rigorous common standards would create equity in American education. However, opponents of these reforms subscribed to an alternative definition of equality, one which placed a premium on pluralism and differential contributions (Airasian, 1987). Such a definition of equality valued diversity of outcomes over equal access to the same content or similar outcomes for all students.

Indeed, some opponents of testing for accountability believed that the student should have a voice in determining the content that they learned (Garrison, 2012; Ohanian, 1999). Content should be individualized to meet the needs and interests of each student (Darling-Hammond, 1994; Ohanian, 1999). Critics of standards, testing and accountability believed that by giving each student a voice in the content that he or she learned, the content would be more relevant to the student and his or her life goals. By allowing students to follow their interests, educators would foster a love of learning. Ohanian (1999) argued that finding joy in learning and

becoming lifelong learners should be considered important outcomes of a quality education (Ohanian, 1999).

However, proponents of rigorous common standards countered that knowledgeable adults were better positioned than children to determine what students should know and be able to do (Finn & Ravitch, 1996). Furthermore, all students needed competency in certain basic skills in order to have access to higher paying employment opportunities (Hanushek, 1989; Rury, 2012).

Opponents of standards not only found value in diversity of content, they also accepted diversity of individual ability. Therefore, they argued, the rigor of content should be individualized to the ability level of each student. Opponents argued that using standards to equalize education for children of different racial, ethnic and socio-economic groups missed the mark, since there was greater variation in ability within any such group than between them (Rothstein, Jacobson & Wilder, 2008). In this respect, they believed that individualizing content to the needs of each student was class and "race" neutral.

According to this perspective, the need to differentiate outcomes by ability was particularly relevant with regard to students with disabilities or who were not proficient in English. Some opponents of standards argued that it was unfair to hold these students to the same standards as their non-disabled or native English speaking peers (Darling-Hammond, 2004; Wood, 2004). For example, NCLB (2002) requirements that English Language Learners (ELL) take reading tests in English ignored the criteria established for students to qualify for ELL services. If they were proficient enough in English to read and write in English, they would not qualify for ELL services. Such requirements were unfair to the students and to their schools.

Furthermore, some critics of standards, testing and accountability protested that standards that were established to prepare students to attend college disregarded the fact that some students

lacked ability or desire to attend college (Ohanian, 1999). Policies that held students and educators accountable for achieving college-ready standards were unfair to these students as well as their schools.

Finally, Darling-Hammond (1994) noted that the argument that all children were capable of learning should not negate the need to differentiate the starting point, the pace of instruction, or the "pathways" by which children achieved the standards. She argued that pacing and starting points should be determined by an assessment of the student, not by the curriculum.

However, it should be noted that promoters of standards, testing and accountability argued that it was unfair *not* to hold educators accountable for teaching all students, regardless of ability, to the same high standards. For too long, many educators expected less from poor and minority children, children with disabilities, and children who were not yet proficient in English. This is what President George W. Bush called "the soft bigotry of low expectations," (Bush, September 2, 2004).

Dehumanization

A theme of both testing debates was the dehumanizing treatment and language used by promoters of objective measures in reference to students, teachers, and education. In the case of the standards, testing and accountability debates, this dehumanization occurred through the application of the language of economics, business, and industry to human beings (McNeil, 1988a; Ohanian, 1999; Sacks, 1999; Walberg, 1998). It occurred through the application of the technical, objective nature of standardization and testing. Finally, it occurred in the reification of human qualities to a numeric score on a standardized test.

Proponents of testing for accountability occasionally used dehumanizing, objective language when referring to human beings and human activities (McNeil, 1988a; Ohanian, 1999; Walberg, 1998). Walberg (1998), for example, referred to education as an industry, and the IQ scores of children as this industry's "raw material," (Walberg, 1998, p. 173). Testing proponents frequently referred to the effect that teachers had on students and student achievement compared to resources invested as teacher "productivity," (Finn, 2002; Finn, Manno, & Vanourek, 2001; Hanushek, 1989, 1994, 2003; Kearns, 1988; Gerstner, 1994; Walberg, 1994, 1998). Opponents of testing noted that proponents referred to teachers as the "objects of production," and test scores and graduates as "final products," (McNeil, 1988a, 1988b; Ohanian, 1999). Education was defined as "the distribution of information." Knowledge was "conveyed" in a "continuous stream" to children by teachers, who were equivalent to assembly line workers. Readers and writers were "consumers and producers of language," (Ohanian, 1999, p. 81).

Ohanian (1999) maintained that the standardization of education also contributed to the objectification of children by treating them as "one undifferentiated mass," (Ohanian, 1999, p. 72-73), "commodities to be regulated (but not paid for) by the government," (Ohanian, 1999, p. 14; see also Sacks, 1999). Standardization was an attempt to establish a uniform curriculum and a uniform child (Ohanian, 1999).

Furthermore, several critics claimed that testing objectified children, educators, schools, as well as the process of teaching and learning by reducing them to numeric scores (Kohn, 2000; Madaus, Russell & Higgins, 2009). Kohn (2000) suggested that testing was based on the simplistic assumption that all things could be measured, and changes in scores represented differences in an "amount" of learning. As such, tests appeared to be objective. Critics of testing suggested that in this technological age, Americans were drawn to reductionist abstractions of

reality (Kohn, 2000; Sacks, 1999). Test scores were an example of such abstractions in that they reduced student learning, student knowledge, or school quality to a number resulting from a paper and pencil test (Kohn, 2000; Madaus, Russell & Higgins, 2009). As with IQ test scores, this reduction of abstract constructs to a number was an example of "reification," (Gould, 1981/ 1996).

Critics argued that this focus on numbers shifted attention away from students and teachers as unique human beings with unique "stories" (Madaus, Russell & Higgins, 2009; Ohanian, 1999). Although tests measured only a small portion of what the child knew or could do, they became a proxy for the worth of the child. Referring to children only in terms of their test scores "delegitimizes them as young human beings," (McNeil, 2000, p. 733). Test scores did not account for "social and collaborative aspects of learning," and students were "subsumed into depersonalized, often meaningless, aggregates," (McNeil, 2000, p. 733; see also Madaus, Russell & Higgins, 2009).This latter point was reminiscent of Dewey's (1922/ 2008a) concern that children lost their individuality when they were classified, sorted, and educated according to their scores on IQ tests.

Critics charged that testing for accountability had a similar effect on schools. As tests were more frequently used to monitor, rate and rank schools, they increasingly came to represent the sole indicator of school quality (Airasian, 1987; McNeil, 2000). As such, high stakes tests allowed testing advocates to "quantify and objectify…schools," (Russell, Madaus, and Higgins 2009, p. 23). Testing for accountability redefined school quality from teaching, resources, and opportunities to learn, to aggregate test scores and the percentage of students who scored above a cut score (Madaus, Russell, & Higgins, 2009; McNeil, 2000; Meier 2004; Wood, 2004). In this

way, the abstract constructs of "teacher ability" and "school quality" and everything that went into them were also reified into numbers.

Tests also gave the appearance that they were objective measures of academic achievement, and were therefore fair. All students were assessed by the same criteria, regardless of their "race" or class.

Meritocracy

Finally, standards, testing and accountability reforms were oriented toward a vision of America as a meritocracy. In a meritocracy, an individual's ability or skill determined social and economic success. As it related to standards, testing and accountability, merit was determined by scores on a standardized test. Schools with high test scores would thrive while those with low scores would collapse. The best educators were rewarded with promotions, raises or bonuses, while the worst educators were fired. Furthermore, promoters of standards, testing and accountability frequently called for minimum test scores to certify that individual students had met the learning requirements of a given grade level, and were prepared to move on to the next (Finn, et al., 2001; Kearns, 1988; Kearns & Doyle, 1989; Walberg, 1998). Such tests were used to hold both students and educators accountable for student learning, and to ensure that both students and educators were motivated to work hard.

Tests that were used to make high stakes decisions such as whether students were promoted, graduated, or were accepted into college became *de facto* economic gatekeepers (Sacks, 1999). Critics charged that such uses of tests unnecessarily and irreparably harmed the future economic opportunities of students who lacked the ability or motivation to do well on the test, or who simply did not test well, by denying them a certification, such as a high school

diploma. This was particularly harmful when test scores were viewed as immutable or infallible (Heubert & Hauser, 1999). Furthermore, because of the correlation between standardized test scores and both "race" and SES, achievement tests joined a long line of tests that effectively helped to maintain the "race" and class based caste system that existed in the United States. However, as discussed earlier, bad schools attended by mostly poor and minority students also served to limit access to a university education and the higher paying occupations (Hanushek, 1989; Kearns, 1988; Rury, 2012). The quality of schools that typically serve poor and minority children may have contributed to their low test scores. Whether they created or limited opportunity, the economic gatekeeper function of tests had serious consequences, particularly for poor and minority students.

Contemporary Relevance

As of this writing, the testing debate continues to be relevant and fluid. At present, the debate over national standards in the form of the new Common Core State Standards (CCSS, 2010) continues. In a 2009 speech to the National Governors' Association, U.S. Department of Education Secretary Arne Duncan (2009) praised the push for common national standards that were "higher, clearer, and fewer," (Duncan 2009, p. 6). Other promoters of the CCSS (2010) argue that common national standards will remedy the uneven quality of national education by standardizing curriculum and controlling for the uneven distribution of leadership in the form of local school boards (Toch, 2011).

What follow are a few of the criticisms that have been raised against the CCSS (2010). In a policy brief for the conservative think tank, *The Heritage Foundation*, Burke and Marshall (2010) argued that the CCSS (2010) were a distraction from the real problem of American

schools: unions and teacher pay schedules. They predicted that national standards would result in increased federal control over education, with weakened accountability. Pressure to scale down toward the lowest common denominator in terms of state standards would result in institutionalized mediocrity. Despite these reservations, in a statement reminiscent of the accountability testing debates, they called for rigorous state standards, and aligned assessments that would provide parents with information to make sound choices in quasi-privatized voucher systems. So, they did not reject standards and accountability testing, *per se*. Rather, they rejected what they perceived to be centralized control of the curriculum.

Other critics also claimed that the CCSS (2010) unconstitutionally gave the Federal government control over education, that local governments were better suited to make curricular decisions, and that the cost of implementing the CCSS (2010) would cause an increase in taxes (Truth in American Education, March 2011). Some critics expressed concern that the continued emphasis that the CCSS (2010) placed on reading, writing, and mathematics dismissed the many other ways that students learn (Strauss, November 7, 2013). Others were concerned over the way in which the CCSS (2010) was being rolled out, and were concerned that teachers would be held accountable for student performance on difficult tests for which they had not had adequate time to prepare their students (Hammond, May 12, 2014; Layton, June 10, 2014; Resmovits, April 30, 2013).

In addition, there has been a controversy over the adoption of value added assessments, particularly the movement to evaluate teachers based upon value added student growth measures. Hanushek (2014), citing his own research (Hanushek & Ramond, 2005) argued that past accountability systems had improved student achievement. Value added accountability systems accounted for forces outside of the teacher's control that impacted student learning. Value added

models eliminated the motive for educators to focus undue attention on so-called "bubble kids" (students likely to score near the cut score), and redirected their efforts to students that were likely to score either far above or far below the proficiency cut scores. Finally, Hanushek argued that teachers should be compensated according to their level of impact on student learning, and suggested that value-added growth measures were a fair way to pay teachers based upon their performance.

Criticisms of value added growth measures included the following: 1) most of the state assessments used for evaluating teachers with value added models were not validated for determining teacher effectiveness (Popham, 2010); 2) many factors outside of the teacher's influence affected student learning (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011); 3) multiple sources of information, including value add on test scores, should be used in evaluating teachers (Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, & Shepard, 2010); 4) Teacher value added scores varied from year to year, and from classroom to classroom (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011); 5) the way schools sorted and assigned students biased student growth measures (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011; Koedel & Betts, 2011).

Finally, a quick scan of the above citations indicates that many of the participants throughout the earlier accountability testing debates continue the debate as it moves into the "value add" and CCSS controversies. However, there is at least one notable exception: Diane Ravitch (Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, & Shepard, 2010; Ravaitch, 2010). Ravitch, one of the more vocal proponents of accountability testing from the late 1980s into the early years of the 21st Century, published what amounted to a retraction in 2010. In her book, *The Death and Life of the Great American School System*, she

claimed that she had been wrong to support standards and accountability testing. Since then, she has been prolific in her attacks against both standards and accountability testing (see for example Ravitch, 2010; Ravitch, May 31, 2011; Ravitch, June 30, 2012; Ravitch, April 6, 2014: Ravitch, May 5, 2014a; Ravitch May 5, 2014b; Ravitch).

In addition, U.S. Secretary of Education Arne Duncan, a supporter of rigorous national standards and the use of tests scores to hold educators accountable for student achievement (Duncan, 2009; 2010; December 11, 2011; June 2013; December 2013; January 2014), recently made remarks in his blog (Duncan, August 21, 2014) suggesting that he may be stepping back, somewhat, from his stand on testing. Duncan stated, "I believe that testing issues today are sucking the oxygen out of the room in a lot of schools…Too much testing can rob school buildings of joy, and cause unnecessary stress," (Duncan, August 21, 2014, no page). It should be noted, however, that while this appears to be something of a retreat, he reiterated the need to measure student growth through assessments, and the need to develop new and better forms of assessments. Still, this moderating of his stance, the retraction of Diane Ravitch, the debate over the common core, and the shift of the accountability testing debate to value added testing model demonstrate that the debates over standards and accountability continue to be a relevant, albeit dynamic issue.

Conclusion

Accountability emerged from criticisms of public education that began with *A Nation at Risk* (Commission for Excellence in Education, 1983). Accountability testing reforms were based on the premise that tests offered an objective means to measure student achievement, and that these achievement test scores provided an objective measure of the quality of the educator or

educational system. A common rigorous curriculum based on standards to which the tests were aligned, and a system of rewards and punishments to be meted out based on achievement test scores further rationalized the system. These proposals would centralize control over education, placing this control external to the classroom.

The criticisms of public education, the proposals for using accountability testing to improve public education, and the suspicions over the motives behind these proposals led critics of testing for accountability to engage proponents in policy debates. These debates tended to center around certain themes, which included the following: the quality of the evidence used to claim that schools were failing; varying definitions of "opportunity to learn" and its effects on educational equity; teacher empowerment and locus of control; the technological/ production/ business orientation of standards based reforms; the proper goals of education in the United States; the distortion of educational goals and experiences resulting from testing and accountability; dehumanizing language; ulterior motives of proponents and critics; and testing as an economic gatekeeper in a meritocracy. These themes will be compared to, and possibly synthesized with the themes identified earlier that emerged from the intelligence testing debates in order to identify the enduring themes of this historical/ comparative analysis. These enduring themes and their implications will be analyzed and discussed in the next and final chapter.

Chapter 6: Enduring Themes and Recommendations

Intelligence tests were designed to measure *ability*, or intelligence. In general, these tests were believed to measure something innate in the individual. Tests used for accountability, however, were designed to measure *achievement*, or what the individual had *learned*. Achievement tests used for accountability purposes were also used to measure the effectiveness of teachers and schools at teaching students. Ability and achievement tests were intended to be used in very different ways. Yet, in the analysis of the respective testing debates, several related themes have emerged that were common to both eras. These included the following: 1) merit; 2) "race," class and educational equity; 3) the meaning of democracy; 4) the fundamental purpose of public education and desired educational experiences in the United States; 5) the role of science and ideology in policy making; and 6) the tendency to oversimplify. The purpose of this chapter is to explore these themes, and to use them to provide guidance to lawmakers as they deliberate future policy decisions concerning educational testing in the United States.

Merit

The United States was founded on the assumption of equality and advancement based upon merit (Kett, 2013). Kett (2013) defined merit as "a quality deserving reward," (p. 10). He differentiated between essential merit and institutional merit. Essential merit was the "inner quality" (p. 3) that resulted in one's "visible and notable achievements…and performances" (p. 2). Essential merit was based on a subjective measure of an individual's inner quality that manifested itself in notable achievement. Examples of essential merit were "intelligence" and "character." An example of an achievement that could be considered to be the outward manifestation of essential merit included leading an army to a victory on the battle field.

Institutional merit was "the acquisition of knowledge of the sort that may be assessed by written examinations," (p. 6). Earned degrees, achieved honors, published articles as well as scores on achievement and intelligence tests were indicators of institutional merit. Institutional merit relied on some sort of external achievement or ability ranking based upon an objective measure.

Although intelligence test scores represented a type of institutional merit, promoters of intelligence tests believed that they had isolated and could measure essential merit in the form of innate intelligence. Intelligence determined an individual's capacity to think abstractly, to learn, to solve complex problems, and to lead (Bagley, 1922a; Goddard, 1922; Haggerty, et al., 1920; Jensen, 1969; Terman, 1919, 1922b). These traits would enable the more intelligent to achieve economic and social success (Goddard, 1922; Herrnstein & Murray, 1994; Duncan 1968; Terman, 1919). To be fair, although IQ testers believed that intelligence was passed from parent to child through "natural endowment," they acknowledged that "intelligence" could appear in an individual of any "race," gender, or social class (Goddard, 1922; Terman, 1919; 1922a; Whipple, 1922). Intelligence tests were believed to be an objective tool for identifying merit no matter where it was to be found.

It is important to note that while these early intelligence testers believed that people at all levels of intelligence could learn (Goddard, 1922; Terman, 1922a), they frequently wrote of those with below average intelligence in terms of limitations. An individual's IQ score indicated the range of professions at which he or she could reasonably expect to succeed (Goddard, 1922; Terman, 1919; Thorndike, 1919). In addition, promoters of intelligence tests dismissed the possibility that factors such as motivation or persistence could improve intelligence (Goddard, 1922; Heckman and Kautz, 2013; Heckman, Pinto, and Salveyev, 2013; Lippmann, 1922c; Rury, 1995). Furthermore, they dismissed the possible influence that the environment, including such

factors as poverty, parenting style, nutrition, or the quality of schooling, could have on intelligence (Brigham, 1923; Goddard, 1922; Herrnstein & Murray, 1995; Jensen, 1969; Terman as cited in Lippmann, 1922e and Lippmann, 1922d; Terman, 1919, 1922a; Whipple, 1922). To the promoters of intelligence testing, ability was predetermined at birth, and therefore out of the individual's control (Brigham, 1923; Goddard, 1922; Herrnstein & Murray, 1995; Jensen, 1969; Terman, 1919, 1922a; Whipple, 1922).

Although advocates of accountability tests believed that their tests measured learning, like intelligence testers they too believed that their tests measured a quality deserving of reward (Finn, et al., 2001; Hanushek, 1994, 2005; Kearns, 1988; Kearns & Doyle, 1989; Ravitch, 1993, 1996). The promoters of testing for accountability believed that they were measuring *achieved* merit. Like the promoters of intelligence testing, the advocates of testing for accountability also dismissed the influence that environmental factors, such as poverty, nutrition, and parenting style, had on academic achievement. However, it was not so much that they denied that such factors might impact achievement, but feared these factors would provide convenient excuses for lack of adequate progress (Hanushek, 2005; Bush, September 2, 2004; Walberg 2001). Instead, they stressed the importance of teacher quality and efforts on improving achievement (Finn, 2002; Finn, et al., 2001; Hanushek, 1994, 2003; Hanushek & Raymond, 2005; Walberg, 1998).

Achievement tests used to hold educators accountable measured the merit of the school. The threat of accountability sanctions pressured schools to teach the tested curriculum. Therefore, the tests defined merit not only for educators and schools, but for students, as well.

Time spent on educative activities was a zero-sum game: time spent on one activity was unavailable for other activities (Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008). Tests with high stakes attached were intended to motivate educators to dedicate increased time to

tested content, reducing the time available for other content or activities (Bracey, 1996; McNeil, 2000, 2004; Kohn, 2000; Madaus, Russell & Higgins, 2009; Ohanian, 1999; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). Furthermore, even the structure of test items determined what skills were valued (Darling-Hammond, 1990; Kohn, 2000; Madaus, Russell & Higgins, 2009; Ohanian, 1999; Sacks, 1999). For example, short answer or multiple-choice questions emphasized an ability to memorize facts, whereas performance events required the ability to use knowledge to solve real-world problems. Because the tests defined what content and skills the teacher taught, and because a diploma represented "merit," the test effectively defined merit.

Critics of intelligence tests as well as critics and proponents of accountability tests tended to argue that merit was not primarily inherent. Nor was merit limited to that which could easily be measured on a test (Bracey, 1987; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008). Critics of testing did not argue against the importance of problem solving ability, the capacity for abstract thought, or the acquisition of specific foundational knowledge and skills that would provide students with the preparation necessary to entry into the higher paying professions an option. However, other traits, such as a sense of humor, creativity, perseverance, compassion, assertiveness, kindness, honesty, motivation, and moral values also contributed to success (Bagley, 1922a; Bracey, 2001; Heckman and Kautz, 2013; Heckman, Pinto, and Salveyev, 2013; Ohanian, 1999; Rothstein, 2008; Rothstein, Jacobson, and Wilder 2008; Rury 1995). These traits were typically not measured by IQ tests or tests for accountability. In addition, minority cultures in the population may have placed a high value on certain traits, traits that in their culture were "deserving of reward," (Dewey, 1922b/ 2008). Because these traits were not valued highly by the dominant culture, they were either not included in the tests, or the tests were designed in ways

that caused these traits to become disadvantageous to test performance (Klineberg, 1928; Dewey, 1922b/ 2008).

In the United States, the rewards for institutional "merit" could be significant. Certifications, diplomas, degrees, completion of programs of study, and test scores placement or professional exams determined who was accepted into the professions, and therefore who enjoyed economic and social rewards. By determining who had access to these rewards, indicators of institutional merit also determined who was denied access to them (Tilly, 1998; Sacks, 1999).

Institutional merit, including achievement test scores, or diplomas, was, at least on the surface, an objective means of determining merit. However, the mechanism by which society established institutional merit had very high stakes for individuals and for groups. High school diplomas determined whether individuals were qualified to enter into college. College degrees and certification exams determined whether individuals were qualified to enter into certain professions. Student test scores used to establish adequate yearly progress could be used to make employment decisions for educators or to determine which schools received additional resources. Therefore, the method for determining who has merit in the future must be established thoughtfully, and in a way that is fair to the most people.

Although tests can be a source of objective information about an individual's learning or qualifications to perform certain types of tasks, the use of tests as a primary indicator of merit may have negative consequences, particularly for subgroups of the population. In particular, the appearance of objectivity in measuring merit may disguise the possibility that the tests are used in ways that disproportionally harm members of subgroups of the population. These possible

negative consequences for class and "racial" groups help to define the next theme that emerged from this analysis, "race, class and educational equity."

Race, class and educational equity

The theme of "race, class, and educational equity" relates to the way that test scores as measures of merit were used to either create opportunities or to exclude lower status groups from valued resources. This theme was common to both testing debates. Both IQ tests and tests used for accountability were promoted as a means for increasing social and economic opportunities for poor and minority children, as well as condemned as tools that reinforced a "race" and class based caste system.

Tilly (1998) argued that large differences in access to valued resources between different groups were related to categorical differences rather than differences in merit. Categories were classifications of human groups. These categories tended to be paired, such as minority/ non-minority or male/ female, and the categorical pairs tended to be of unequal status. Such inequalities arose when those who controlled access to "value producing resources…inadvertently or otherwise" created systems that excluded and controlled members of the lower status group (Tilly, 1998, p. 8). Furthermore, inequalities that observers often attributed to individual differences were actually caused by structured inequality based upon such paired categories.

Structured inequalities were established through "exploitation" and "opportunity hoarding," (Tilly, 1998). Exploitation occurred when high status, high power groups created structures that enabled them to increase their own access to valued resources by restricting this access to the lower status groups that produced or increased the availability of the resource. An

obvious example of exploitation was slavery. Opportunity hoarding occurred when monopoly access to valued resources was created by a group's mode of operation. An example of opportunity hoarding was the literacy requirements for voting in the Jim Crow South.

Critics of both intelligence tests and tests for accountability charged that the actual as well as proposed uses of these tests were for purposes that amounted to opportunity hoarding (Bagley, 1922a, 1922b, 1923; Lippmann, 1922f; Sacks, 1999). As discussed elsewhere, critics charged that both intelligence and accountability tests and related sanctions had a disparate impact on poor and minority individuals (Bagley, 1925, Chapter VI; Bond, 1924a, 1924b, 1934/ 1966; Darling-Hammond, 1994, 2004; Kohn, 2000; Ohanian, 1999; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004). During the World War I Army testing program, intelligence tests were, in part, intended to place soldiers in jobs. As Tyack (1974) stated, this may have determined whether an individual was given an office job in Washington, or sent to the trenches of France. Intelligence tests were used to justify tracking students into college preparatory or vocational classes (Terman, 1919). Scores on the SAT, which originally was an adaptation of the Army Alpha and until recently was considered to be an "aptitude" test rather than an achievement test, are used as criteria for determining whether students were accepted into colleges or universities (Lemann, 1999/ 2000; Sacks, 1999). As discussed in Chapter 4, intelligence tests were also used to rationalize ending welfare, compensatory education and affirmative action (Herrnstein & Murray, 1994; Jensen, 1969).

Almost from their inception, intelligence tests were known to be related to "race" and social class (Bagley, 1925, Chapter VI; Bond, 1924a, 1924b; Brigham, 1923; Goddard, 1922; Terman, 1916, 1919, 1922b;). Even those who desired to use the new tests to create opportunities for intellectually gifted but marginalized people believed that there was a hierarchy of "races"

that had evolved through natural selection, and that group averages on intelligence tests reflected these differences (Goddard, 1922; Terman, 1919, 1922a; Whipple, 1922). Average group intelligence test scores tended to reinforce these beliefs. Advocates of intelligence testing did not question the circumstantial evidence that supported their *a priori* held beliefs (Brigham, 1923). Furthermore, because the tests were purported to open opportunities for any individual with merit, they had the effect of providing a rationale for the exclusion of large numbers of poor and minority people from access to economic and social rewards.

Just as intelligence testers promoted IQ tests as a means for opening opportunities for marginalized groups, one of the stated goals of testing for accountability was to improve schools for all children, including poor and minority students (Ravitch, 1996, p. 134; see also Kearns, 1988; Ravitch, 1993; Smith & O'Day, 1991). Tests aligned to rigorous standards would force schools to teach all children to these standards. Failure to educate all children to these standards could result in school closures and the removal of incompetent educators from the classroom (Wallberg, 2001). In some cases, tests and accountability would provide poor and minority children a means of escape from failing schools (DeBray-Pellot & McGuinn, 2009; Gamoran, 2007b). Indeed, some of the staunchest advocates of testing for accountability were members of minority groups that had experienced some of the nation's worst schools (Archer, 2006; Loveless 2007; Reid, 2005; Salzman, 2006). In essence, they believed that holding educators accountable for test scores would force educators to dismantle some of the structures that excluded poor and minority children from access to good schools, and, as a result, valued resources.

Alternatively, critics of testing for accountability argued that because of the environmental challenges facing poor and minority school children, these children were more likely to score poorly on standardized tests (Darling-Hammond, 2000; Karp, 2004; Kohn, 2000;

Koretz, 2008; Madaus, Russell & Higgins, 2009; Ohanian, 1999, 2003; Sacks, 1999; Rothstein, 2004, 2008). As such, sanctions associated with testing for accountability undermined schools serving higher proportions of poor and minority students. They argued that sanctions could negatively impact education for these children in a number of ways. Sanctions that called for supplementary educational services such as tutoring, or that required schools to pay tuition and transportation costs for students to attend non-failing schools placed additional financial burdens on already under-resourced schools (Darling-Hammond, 2004; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). Sanctions could discourage highly qualified educators from working in under resourced urban and rural school districts serving high proportions of poor and minority school students (Darling-Hammond, 1994, 2004; Kohn, 2000; Ohanian, 1999; Sacks, 1999; Wood, 2004).

Additionally, educators in schools under greater pressure to improve were more likely to behave in ways that narrowed educational experiences, especially for poor and minority students (Darling-Hammond, 2000; Madaus, Russell, & Higgins, 2009; McNeil, 2000; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). Such behaviors included restricting non-tested content and activities (Bracey, 1996; Darling-Hammond, 1990; Kohn, 2000; Madaus, Russell & Higgins, 2009; McNeil, 2000, 2004; Ohanian, 1999; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004), teaching to the test (Bracey, 1987; Darling-Hammond, 1994; Madaus, Russell & Higgins, 2009; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004), focusing efforts on students near proficiency cut scores at the expense of those farther away (Koretz, 2008; Madaus, Russell & Higgins, 2009; Rothstein, 2008), and outright cheating (Koretz, 2008; Rothstein, Jacobson & Wilder, 2008). Finally, the standards, testing and accountability reforms allowed policy makers to claim that they were

doing something to improve schools for all children, without addressing the often dramatic differences in resources provided to public schools serving poor and minority students as compared to middle and upper class non-minority students (Darling-Hammond, 2004; Karp, 2004; Kohn, 2000; Madaus, Russell & Higgins, 2009; Meier 2004; Ohanian, 1999; Sacks, 1999; Wood, 2004).

Furthermore, some critics claimed that testing for accountability originated from a desire to serve the business and economic sector (Sacks, 1999). Tests would ensure that workers had the minimum knowledge and skills necessary to serve the needs of employers (Bracey, 1996; Ohanian, 1999). Furthermore, they would create a glut of skilled labor, which would drive down the cost of wages (Ohanian, 1999, 2000, 2003). Assuming these critics were correct, such a policy would be part of a larger mechanism for exploiting the lower status groups.

In addition to exploitation and opportunity hoarding, critics from both testing debates argued that the use of numbers and language to objectify humans were ways that testers from both eras maintained inequality (Bagley, 1922a, 1922b, 1925; Bond, 1924a, 1924b, 1934/ 1966; Dewey, 1922/ 2008b; Gould, 1981/ 1996; Jackson & Weidman, 2004/ 2006; Lippmann, 1922d; McNeil, 1988a, 1988b; Ohanian, 1999; Sacks, 1999). As discussed in previous chapters, promoters of intelligence testing as well as testing for accountability spoke of humans or human institutions in terms of scores on intelligence or achievement tests (Kohn, 2000; Lippmann, 1922a, 1922c; Madaus, Russell & Higgins, 2009; McNeil, 2000). Using a number to represent an abstraction was an example of the logical fallacy of reification (Gould, 1981/ 1996). Furthermore, it placed a numeric value on humans, allowing them to be ranked by inherent worth (Kohn, 2000; Lippmann, 1922a, 1922c; Madaus, Russell & Higgins, 2009). As discussed in

chapters 3 and 5, the assignation of a numeric value to humans was one way that they were treated as objects (Madaus, Russell & Higgins, 2009; McNeil, 2000).

Similarly, proponents of both intelligence and accountability testing used the language of objects when referring to humans or human traits (Finn, 2002; Finn, Manno, & Vanourek, 2001; Goddard, 1922; Haggerty, et al., 1920; Hanushek, 1989, 1994, 2003; Kearns, 1988; Gerstner, 1994; Spearman, 1904; Terman, 1919, 1920; Thorndike, 1919; Walberg, 1994, 1998; Whipple, 1922). Words and analogies borrowed from business, manufacturing, or the natural sciences were used to describe students, teachers, and schooling (Finn, 2002; Finn, Manno, & Vanourek, 2001; Gerstner, 1994; Goddard, 1922; Haggerty, et al., 1920; Hanushek, 1989, 1994, 2003; Kearns, 1988; McNeil, 1988a, 1988b; Ohanian 1999, 2000; Spearman, 1904; Terman, 1919, 1920; Thorndike, 1919; Walberg, 1994, 1998; Whipple, 1922). As when they referred to people as their score on a test, this practice also effectively treated humans as though they were objects.

The reference to humans as though they were objects was important for psychological reasons. It was difficult for people to treat other humans inhumanely, so long as the other was thought of as fully human. However, when perceived to be less than human or not human, inhumane treatment became easier. As Freire (1990) pointed out, "…the more the oppressors control the oppressed, the more they change them into inanimate 'things,'" (Freire, 1990, p. 45). Using objective language psychologically converted the low status group into resources, commodities to be exploited. Freire wrote, "The oppressor consciousness tends to transform everything surrounding it into an object of its domination. The earth, property, production, the creations of men, men themselves, time—everything is reduced to the status of objects at its disposal, (Freire, 1990, p. 44).

To be sure, promoters of testing in both eras had a desire to present testing as an objective measure of merit. As mentioned above, Kett (2013) suggested that institutional merit was, or at least gave the impression of being, a more objective representation of merit than was essential merit. However, the appearance of objectivity served to maintain inequality. By giving the appearance of objectivity, it provided the high status group with the rationale for oppressive practices, and restricted the ability of low status groups to argue against these practices. There was no better example of this than the way IQ scores were used to oppress marginalized groups in this and other countries.

Science in the service of hegemony created myths that helped to sustain unequal power relationships (Friere, 1990). Intelligence tests promoted the myth that the poor and certain "racial" and ethnic groups were less intelligent than middle and upper class whites, and therefore deserved their lower status. The myth was particularly effective when it was internalized by members of the lower status group. As Freire (1990) wrote, "Self-depreciation is another characteristic of the oppressed which derives from their internalization of the opinion the oppressors hold of them. So often they hear that they are good for nothing, know nothing and are incapable of learning anything—that they are sick, lazy and unproductive—that in the end they become convinced of their own unfitness," (Freire, 1990, p. 49).

Tests can provide valuable information about the level of student achievement, the identification of possible learning problems, and the effectiveness of educators. Furthermore, accountability tests used for NCLB (2002) have exposed the achievement of subgroups of the population, thereby holding educators accountable for educating all children (Rury, 2012). However, tests may also disproportionately harm poor and minority children. Student achievement, including achievement on test scores, is related to the educational attainment of

their parents (Coleman, et al., 1966; Rury & Akaba, 2014). When test scores are used as a primary source of information for determining opportunities for future educational attainment and access to valued resources, tests can become part of a vicious cycle that maintains the existing social hierarchy. Poor and minority children are more likely to have lower test scores, resulting in fewer opportunities for educational attainment among these populations. As these children become adults, their lower educational attainment may translate to lower test scores and educational attainment for their children, as well. Similarly, schools serving poor and minority children are more likely to have lower test scores, and are therefore more likely to face sanctions. Sanctions or the threat of sanctions that restrict the capacity of these schools to provide a quality education may result in lower test scores, leading to even more sanctions. These lower scores may also limit future educational attainment for these students, contributing to the lower test scores and educational attainment of their children.

The Meaning of Democracy

The themes of "merit" and "race, class and educational equity" were intertwined with how participants in the debates defined democracy and equality. For several of the participants in these debates, equality was an integral part of how they defined democracy. Indeed, what often differentiated the definitions of democracy for the various participants in these debates was how they chose to answer the question, "equality of what?"

As discussed above, intelligence testers did not believe that people were born equal. Rather, those born with greater intelligence were better suited to lead than were individuals born with less intelligence (Goddard, 1922; Terman, 1919, 1922a; Whipple, 1922). As such, promoters of intelligence testing defined democracy as "equity of opportunity" to lead based

upon personal merit (Whipple, 1922, p. 602; see also Terman, 1919, 1922b). That is, individuals should not be artificially limited by "race," class, sex, or family connections. Instead, any individual should have the freedom to achieve success or to become a member of the new intellectual aristocracy based solely on innate personal merit. Goddard (1922) went so far as to suggest that a minimum IQ score should be a criterion for voting.

By contrast, critics of intelligence testing such as Bagley (1922a) argued that democracy meant collective rule. Such a system of governance required universal education to ensure that an educated populace could make sound political decisions. This definition was similar to that held by the intelligence test promoters, in that these critics also argued for equality of opportunity to lead and make decisions based on merit. However, this perspective defined merit more broadly to include such traits as character, and attributed a larger proportion of merit to learning.

Dewey (1922/ 2008a; 1922/ 2008b) held yet another perspective on democracy. His definition of democracy was based upon his distinction between individualism and individuality. Individualism was the right of the individual to pursue personal success which was dependent upon a socially sanctioned definition of "merit." By contrast, Dewey defined individuality as the right to pursue individualized interests based upon personal values. In this definition, merit was defined by each individual and this definition was grounded in cultural values. For Dewey, democracy was equality of status among unique individuals.

Like the promoters of intelligence testing, advocates of testing for accountability defined equality as equality of opportunity to achieve merit and to benefit from the rewards that accompany achieved merit (Gamoran, 2007b; Kearns, 1988; Ravitch, 1996; Rury, 2012). This, in turn, depended upon equality of access to a rigorous curriculum and instruction, as well as

equality of certain minimum educational outcomes, regardless of class, "race" or ethnicity (Gamoran, 2007b; 1996; Kearns, 1988; Ravitch, 1993; Rury, 2012; Smith & O'Day, 1991). Similar to that proposed by IQ test promoters, this definition of equality consisted of a universal, socially sanctioned, externally determined, definition of merit (Ravitch, 1996). However, like the critics of intelligence testing, promoters of testing for accountability emphasized that merit was largely composed of learned knowledge and skills.

Finally, similar to Dewey's definition of individuality, at least one critic of testing for accountability defined equality as the equal status among unique individuals living in a pluralistic society (Airasian, 1987). This definition emphasized the equal but different contributions of different people.

These different conceptions of democracy and equality provide a link between the themes of merit, "race" class and educational equity, and the fundamental purpose of education in the United States. The definition of democracy and equality to which one subscribed was related to who defined merit, and whether test determined merit was used to restrict access to opportunities and resources. In addition, the difference between the conceptions of individualism and individuality as they related to Dewey's (1922/ 2008a; 1922/ 2008b) definition of democracy had direct implications for the purpose of schooling and desired educational experiences for students in the United States. The theme of the fundamental purpose of education and desired educational experiences is the next subject for this analysis.

The Fundamental Purpose of Education in the United States

The themes identified thus far in this analysis have exposed the fact that the United States has never arrived at a shared belief regarding the fundamental purpose of public education, or the

kinds of educational experiences we want for our children. Is the fundamental purpose of education in the United States to improve the overall economy, or is it to help each individual fully develop his or her unique interests and talents? Is the purpose of public education to prepare students to become economically viable adults, or to prepare students for the life of the mind? Is the purpose to provide a diverse population with a common, American culture, or should public education honor the diversity of America? The perspective that participants of the debates took regarding the fundamental purpose of education in the United States appeared to both influence and be influenced by their beliefs regarding the proper role of educational testing.

Advocates of testing in both debates tended to express the belief that education should serve the economic and business interests of the country, and to prepare students to be economically viable participants in the economy (Callahan 1962; Commission for Excellence in Education 1983; Finn, et al., 2001; Goddard, 1922; Haggerty, et al., 1920; Hanushek, 1989, 2005; Kearns, 1988; Kearns & Doyle, 1989 Terman, 1919). Tests were used to direct students into professions for which they were qualified, and for which there was a business or societal need (Callahan 1962; Goddard, 1922; Haggerty, et al., 1920; Kearns and Doyle, 1989; Terman, 1919, 1922a, 1922b; Whipple, 1922). Education was to prepare future workers to be innovative, good workers, contribute to the overall economy, and be economically self-sufficient.

While IQ testers tended to see testing as a mechanism for social mobility (Goddard, 1922; Terman, 1919; Whipple, 1922), advocates of accountability testing endorsed the belief that education served this function. Promoters of accountability testing argued that all students should be provided access to high quality instruction in a rigorous college preparatory curriculum (Finn, Kanstoroom, Rothstein, & Honig, 2001; Kearns, 1988; Ravitch, 1993, 1996). This goal was particularly important for poor and minority students, who historically had been denied access to

the type of instruction necessary to prepare them for higher education and subsequent entry into high paying professions.

Critics of testing had slightly more diverse opinions on the purpose of public education in the United States. Some, like Bagley (1923, 1925) or Bracey (1996), expressed the belief that public schools should prepare students for "the life of the mind." Such an education would prepare all students to benefit from and to take pleasure in the literary, artistic, and intellectual achievements of the human race. In addition, Bagley (1922a, 1925), argued that public education should provide students from all walks of life with a common cultural experience. This common experience would bond together a diverse population, thereby acting as a social leveler (Bagley, 1925). Finally, Bagley (1922a) also emphasized the need to prepare individuals living in a democracy to be good citizens by providing them with the knowledge and skills necessary to make sound political decisions. Bagley wrote, "With no fear of contradition [*sic*], I can affirm that the safest guarantee of sincere and responsible leadership lies in a level of informed intelligence among the rank and file that will enable the common man to choose his leaders wisely, scrutinize their programs with sagacity and, in the pungent slang of the day, tell them 'where to get off' when they go wrong," (Bagley, 1922a, p. 380).

Some critics of testing from both debates promoted the belief that public schools should be structured to allow students to pursue individual goals and interests (Darling-Hammond, 1994; Garrison, 2012; Ohanian, 1999). Building educative experiences on the personal experiences of the student (Dewey, 1938), allowing them to develop unique talents (Dewey, 1922/ 2008b), and following the interests of the students (Ohanian, 1999) would make education personally meaningful, and therefore foster a love of learning (Ohanian, 1999; Dewey, 1938). Students that had no interest in or ability to pursue higher education should not be required to

learn content designed to prepare them for entrance into a college or university (Ohanian, 1999). Furthermore, educational experiences should be generated by the student in consultation with his or her teacher (Ohanian, 1999), or individualized by the teacher to build on the personal experiences of the student (Dewey, 1938). These perspectives required faith in the student's and/or teacher's abilities to identify a program of study that was in the student's best interest. A standardized, externally imposed curriculum that was policed with aligned tests was contrary to this view of education.

The beliefs about the purpose of education in the United States had implications for the educational experiences of students. Intelligence testers were determinists (Lippmann, 1922f; Bagley, 1922a, 1922b). They believed that the ability to think and learn was determined at birth, and that these native abilities determined the profession for which the individual was best suited, the strength and quality of the individual's character, and the individual's chances for life success (Goddard, 1922; Terman, 1919; Whipple, 1922). Not only did IQ determine the individual's future prospects, but the testers proposed using IQ scores to determine the type of education an individual received, and the type of profession toward which he or she should be directed. Therefore, the educational experiences were out of the individual student's control. Rather, the student was placed in an educational track based upon his or her ability. The content of the track was selected so as to prepare the student for a vocation, profession, or further education. Indeed, at least one advocate of intelligence testing believed that the only way to reach children with low intelligence was through associative learning (Jensen, 1969).

Similarly, the promoters of testing for accountability also advocated for a kind of educational determinism, albeit a very different determinism from that promoted by the intelligence testers. Many of the practices advocated by promoters of testing for accountability

seemed to endorse the belief that environmental conditions determined human behavior, and could be manipulated to shape teaching and learning in the classroom. The promoters of testing for accountability advocated the identification of educational objectives in the form of standards (Bracey, 1987; Finn, et al., 2001; Ravitch, 1993, 1996; Smith & O'Day, 1991). These standards were selected to fill a perceived societal need (Finn, et al., 2001; Hanushek, 1994, 2005; Kearns, 1988; Kearns & Doyle, 1989; Ravitch, 1993, 1996). Furthermore, standards were selected and written so as to be measurable (Finn, et al., 2001). Tests and accountability systems were used to force the entire system to align with the standards, including teacher preparation, professional development, and classroom instruction (Ravitch, 1996; Smith & O'Day, 1991). The learning environment was controlled by imposing rewards and punishment designed to motivate teachers to teach the standards and to improve instruction (Finn, 2002; Finn, et al., 2001; Hanushek, 1989, 1994, 2003; Hanushek & Raymond, 2005; Walberg, 1998).

Furthermore, critics suggested that testing for accountability motivated educators to teach in ways that mirrored standardized multiple choice tests. According to this argument, skills and content were fragmented, and dissociated from the way that they would be experienced by the child in real life settings (Bracey, 1987; Kohn, 2001; Madaus, Russell & Higgins, 2009).

This identification of learning objectives, fragmentation of the knowledge and skills into learnable chunks, and the manipulation of the environment through the systematic application of rewards and punishments suggests a behaviorist approach to education (Bracey, 1987). Indeed, this behaviorist approach was acknowledged by Finn and Kanstoroom (Finn, et al., 2001) when they suggested that "a behaviorist aspect of this reform strategy discomfits some educators, yet it is the way most of the world works," (p. 133). While it is probably unfair to characterize the promoters of both intelligence and accountability testing as behaviorist, the behaviorist

techniques characteristic of the respective testing movements aligned with a technological/ production orientation similar to the scientific management influenced education described by Callahan (1962).

Finally, while generally associated with what Tyack (1974) described as pedagogical progressivism (Finn, et al., 2001; Finn & Ravitch, 1996; Tyack, 1974), critics of testing tended to vary from each other in their beliefs about the types of educational experiences that they valued for children. One common theme, however, was that education should not be limited in scope to tested skills or content (Bagley, 1923, 1925; Bracey, 1996; Dewey, 1922/ 2008b; Kohn, 2000; Madaus, Russell & Higgins, 2009; McNeil, 2000, 2004; Ohanian, 1999; Rothstein, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). Critics associated with the pedagogical progressives from the different debates would likely give the child and/ or teacher more direct control over educational experiences, the content would be taught in ways that crossed traditional disciplinary lines, and it is more likely that content would be learned solving authentic or real life problems (Dewey, 1938; Ohanian, 1999). Beyond that, it is difficult to determine, based upon the sampled works, what types of educational experiences were valued by other critics of test based reforms.

The Role of Science and Ideology in Policy Making

Political ideology was influential in both the intelligence testing debates and the testing for accountability debates. In both debates, the promoters of testing took what might be considered a social Darwinist perspective concerning the role of government in the affairs of citizens (Apple, 2001; Brigham, 1923; Finn, 2002; Finn, et al., 2001; Harvey, 2005; Herrnstein & Murray, 1994; Jensen, 1969; Kearns, 1988; Kearns & Doyle, 1989; Ravitch, 1994; Walberg,

1994, 1998). In such a perspective, individuals and organizations succeeded or failed, thrived or struggled, lived or died according to personal merit. Competition led to innovation and progress. When the weak failed or died, it made the population and society stronger (Minton, 1987/ 1990; Spencer 1864).

The promoters of intelligence testing frequently expressed concern over the "dysgenic effects" of allowing people of lower intelligence to enter the country or reproduce (Brigham, 1923; Herrnstein & Murray, 1994; Jensen, 1969). Furthermore, since they believed that poverty was caused by a lack of intelligence, government interventions intended to help the poor would only exacerbate these problems by allowing the unintelligent poor to survive long enough to reproduce (Herrnstein & Murray, 1994; Jensen, 1969). As such, people should be allowed to make their own way in the world, to thrive or die without government intervention.

Likewise, advocates of testing for accountability took a similar view toward educators and schools. Those educators and schools that produced high achievement test scores should be allowed to thrive, while those that produced poor achievement test scores would be allowed to fail (Finn, et al., 2001; Kearns, 1988). Some critics recommended that this process would be facilitated through market style competition, and therefore recommended incorporating such competition into school accountability schemes (Finn, 2002; Finn, et al., 2001; Kearns, 1988; Kearns & Doyle, 1989; Ravitch, 1994; Walberg, 1994, 1998). Not only would such accountability mechanisms improve education overall, they also improved educational opportunities for children whose neighborhood schools were failing (DeBray-Pellot & McGuinn, 2009; Finn, et al., 2001; Gamoran, 2007b).

This was not to suggest that advocates of testing refused all government interventions. Promoters of intelligence testing advocated using intelligence tests to rank students and educate

them to fill a needed niche (Goddard, 1922; Terman, 1919, 1922a, 1922b; Whipple, 1922). Similarly, advocates of testing for accountability desired to use testing and government intervention to determine the knowledge and skills that served the economic and business interests of the country (Hanushek, 1989; Kearns, 1988; Kearns & Doyle, 1989), and to strengthen good schools while accelerating the demise of poor schools (Apple, 2001; Finn, et al., 2001; Harvey, 2005; Kearns, 1988).

Of particular concern was the influence that ideology had over the respective testing debates, both in terms of interpreting empirical evidence as well as informing policy decisions. With regard to the intelligence testing debate, both promoters and critics of intelligence testing inferred causation from relationships (Brigham, 1923; Heckman, 1995; Herrnstein & Murray, 1994; Lippmann, 1922e; Terman, 1919, 1922b). When evaluating compensatory programs, neither side of the debate conducted cost benefit analyses (Heckman, 1995). Promoters of intelligence testing applied population specific statistics to other populations (Dobzhansky 1973; Gould, 1974/ 1975/ 1999, 1971/ 1975/ 1999, 1981/ 1996; Lewontin, 1970/ 1975/ 1999; Montagu, 1975/ 1999). They demonstrated a lack of understanding of how genes interacted with other genes and with the environment to produce a phenotype (Gould, 1974/ 1975/ 1999, 1981/ 1996; Lewontin, 1970/ 1975/ 1999; Montagu, 1975/ 1999a, 1975/ 1999c). Finally, they relied heavily on likeminded researchers to support their claims (Kamin, 1999; Lane, 1975/ 1999; Ryan, 1999).

Likewise, promoters and critics of testing for accountability often provided *no* evidence to support their beliefs, ignored evidence that contradicted *a priori* held beliefs, or presented one sided arguments that supported these beliefs. For example, Walberg (1998) argued that compensatory programs created a two tiered educational system that stigmatized children and denied them access to the core curriculum, without providing evidence that this was the case. In

the same way, other promoters of testing for accountability argued that market competition would improve schools without providing any evidence to support this claim (Finn, 2002; Finn, Kanstoroom, Rothstein, & Honig, 2001; Kearns, 1988; Kearns & Doyle, 1989; Ravitch, 1994; Walberg, 1994, 1998). Finally, promoters of testing for accountability reported one-sided correlational evidence to show that additional resources did not improve student achievement (Hanushek, 1989, 1994; 2003, 2005; Kearns, 1988; Walberg, 1994).

Similarly, critics of testing for accountability claimed that it was disingenuous to hold all schools accountable when resources were not equitably distributed, but failed to provide evidence that additional resources would improve student achievement (Darling-Hammond, 1994, 2000, 2004; Karp, 2004; Kohn, 2000; Ohanian, 1999, 2003; Rothstein, 2004; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004). Furthermore, critics such as Berliner and Biddle (1995) used conjecture to argue that American students did poorly on tests of international comparison because American parents valued well-rounded students, and therefore American students had limited opportunity to learn tested content.

This theme of allowing ideology to influence the arguments of the debates is related to the final enduring theme, the tendency of both sides of the respective debates to oversimplimplify.

Oversimplification

Both sides of the respective testing debates were guilty of a tendency towards oversimplification, both in terms of their arguments as well as their solutions to educational and social problems. Promoters of intelligence testing dismissed the influence of the environment, experiences, or character traits on intelligence test scores or socioeconomic success (Brigham,

1923; Goddard, 1922; Herrnstein & Murray, 1994; Jensen, 1969; Terman, 1919, 1922, 1922a). They used simple measures of SES as a proxy for the sum of all environmental and educational experiences (Heckman, 1995; Rury, 1995). They assumed that intelligence was genetically related to "race," even when more plausible environmental explanations were available (Brigham, 1923; Goddard, 1922; Herrnstein & Murray, 1994). They assumed that because compensatory programs had a limited and transient effect on IQ scores, they were ineffective (Heckman, 1994; Heckman and Kautz, 2013; Heckman, Pinto, and Salveyev, 2013; Herrnstein & Murray, 1994; Jensen, 1969). Finally, they attributed many social problems to the lack of inherent intelligence among segments of the population (Goddard, 1922; Herrnstein & Murray, 1994; Jensen, 1969).

Critics of intelligence testing made similar mistakes. They inferred causation from a correlation (Bagley, 1925; Lippmann, 1922d). They failed to conduct a cost benefit analysis before endorsing the effectiveness of compensatory education and other social welfare programs (Heckman, 1995). They rejected empirical evidence that contradicted their beliefs (Heckman, 1995). Finally, it was naïve to believe that a single number could summarize the ability or essential merit of an individual (Gould 1981/ 1996).

Critics of public education and promoters of accountability testing during the 1980s and 1990s declared public education as failing based upon test scores (Commission for Excellence in Education, 1983; Hanushek, 1989, 2003; Ravitch, 1999; Walberg, 1998), including scores on voluntary college entrance exams. They rejected the effect of resources as having a positive impact on student achievement, providing only relational evidence to support the claim (Finn, 2002; Hanushek, 1989, 1994, 2003, 2005; Hanushek & Raymond, 2005; Kearns, 1988; Walberg, 1994). They promoted market based reforms without any evidence that they would improve

student achievement or educational equity (Finn, 2002; Finn, et al., 2001; Kearns, 1988; Kearns & Doyle, 1989; Ravitch, 1994; Walberg, 1994, 1998). They promoted standards, testing and accountability reforms without any evidence that this would improve student achievement (Finn, et al., 2001; Smith & O'Day, 1991). They assumed measureable standards to be the only standards worth teaching (Finn, et al., 2001). Finally, they dismissed the role of environmental factors such as poverty or inequitable school funding policies as contributing to poor student achievement and low test scores (Hanushek, 2005; Bush, September 2, 2004; Walberg 2001).

Critics of accountability testing assumed that the purpose of testing for accountability laws such as NCLB (2002) was to cause massive school failure so as to build consensus for school privatization schemes (Karp, 2004; Kohn, 2004b; Wood, 2004). They assumed that additional resources would improve education, without any evidence that this was the case (Darling-Hammond, 1994, 2000, 2004; Karp, 2004; Kohn, 2000; Ohanian, 1999, 2003; Rothstein, 2004; Rothstein, Jacobson & Wilder, 2008; Sacks, 1999; Wood, 2004). Likewise, they assumed that a more equitable distribution of resources would improve educational equity, without any evidence to support this claim (Karp, 2004; Kohn, 2000; Meier 2004; Ohanian, 1999; Rothstein, 2004, 2008; Rothstein, Jacobson & Wilder, 2008; Wood, 2004). Finally, some inferred causes of lower national rankings that were based on conjecture rather than tested hypotheses (Berliner & Biddle, 1995).

Dewey (1938/ 1997) rather famously rejected dualistic thinking. In doing so, he was arguing against oversimplifications. Curriculum should not flow from the whims of the child any more than it should ignore the interests or experiences of the child. Similarly, all testing is neither good, nor bad. Tests can be valuable sources of information that can improve learning for students. However, certain uses of tests can have negative impacts.

Discussion

The preceding analysis has demonstrated that educational testing policies are driven by the assumptions, beliefs, and values orientations of those who write and promote them. These assumptions may reflect beliefs about how children best learn, or even the extent to which certain children *can* learn. They reflect assumptions about the best way to improve instruction in public schools. They reflect values orientations, which may include a desire to improve educational opportunities for poor and minority children, or a desire to protect a monopoly on opportunity for a certain "race," class, or constituency. Testing policies may be driven by values that have little to do with the educational interests of children. For example, testing policy may reflect a desire to maintain or lower taxes for a certain constituency, or to funnel tax dollars to private schools or school management companies. The point is, that educational testing policy creates winners and losers, and, whether unintentionally or by design, the winners may not necessarily be school children.

Testing policies and policy proposals have consequences, whether intended or unintended. Proposals for both intelligence and accountability testing promoted merit systems based upon a type institutional merit that reinforced existing "race" and class based social hierarchies. Both promoted a type of democracy that involved "rule by the best," and the freedom to pursue social and economic success based upon a standardized, socially sanctioned conception of merit. Both promoted a view of education that included a fragmented and compartmentalized view of curriculum, behaviorist instructional methods, and a technological/ production orientation toward efficiently processing children. Finally, both promoted a belief

that the fundamental purpose of schooling was to serve the economic and business interests of the country.

The testing debate has contemporary relevance. The current debate over the adoption of the Common Core State Standards and the new tests that accompany are raising many of the same concerns. The high stakes attached to the new tests make the arguments from the testing debates relevant in that the new high stakes tests could replace the new standards if educators feel pressured to teach to the test. In addition, the debates are in flux, with some former advocates of accountability testing either back-peddling (Duncan, August 21, 2014) or renouncing their former position (Ravitch, 2014).

Educational testing as one source of data can provide valuable information to teachers about student learning and the effectiveness of their instructional strategies. Tests provide educators with one source of information that can help them to diagnose possible learning problems in children. Tests can provide a source of information as to the effectiveness of educators at educating all children in their care. However, when test scores are the sole source of information in evaluating students or schools, when single test scores carry high stakes for students and/ or educators, when test scores become the end goal of education rather than a source of information, testing policies can have negative impacts on students and schools.

As lawmakers contemplate future educational testing laws they need to carefully consider the implications of all proposals before voting. All laws have intended and often unintended consequences. Lawmakers must consider these possible consequences. In addition, lawmakers should carefully consider the possible effects of Campbell's law (Campbell, 1975) on the educational outcomes and experiences for students, when scores on a single test are used to hold

educators accountable. Finally, all laws have winners and losers. Hopefully, lawmakers will pass laws where the winners are the schoolchildren.

Summary

The purpose of this study was to examine the dispositions, assumptions and values orientations of those who promoted testing as compared to critics of testing. This information was intended to be used as a guide for policy makers as they explore future testing policies and the types of educational experiences and valued outcomes that specific policies promote. This study demonstrated that the testing policy and the test design that are adopted determine the beliefs, values, outcomes and experiences that are advanced and endorsed. The purpose was not to evaluate the effectiveness of standards, testing and accountability reform in raising student achievement or closing achievement gaps. Indeed, insofar as accountability measures work when they change the practices of educators to align their teaching with the content tested and the types of problems on the test, a case can be made that an evaluation of the effects of testing for accountability on student achievement will depend on how well the respective state accountability test aligns with the type of assessment used to evaluate the respective state testing programs. How this can be accomplished in a way that is fair, objective, and comparable across states will be left for other researchers to address.

References

Apple, M. (2001). *Educating the "right" Way: Markets, standards, God, and inequality*. New York: Routledge/ Falmer.

Archer, J. (2006). Civil rights groups back NCLB law suit. *Education Week 25*(22), 15-16.

Ayres, L. (1909). *Laggards in our schools: A study of retardation and elimination in city school systems*. New York: Charities Publication Committee.

Ayres, L. (1920). *An index number for state school systems*. New York: The Russell Sage Foundation.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, California: Wadsworth/Thompson.

Bagley, W. (1922a) Educational Determinism; Or Democracy and the I.Q. *School and Society 15* (380), pp. 373-384

Bagley, W. (1922b). Professor Terman's determinism: A rejoinder. *The journal of Educational Research, 6*(5), 371-385.

Bagley, W. (1925). *Determinism in education: A series of papers on the relative influence of inherited and acquired traits in determining intelligence, achievement, and character*. Baltimore, Maryland: Warwick and York, Inc.

Baker, E. (1994). Researchers and assessment policy development—A cautionary tale. *American Journal of Education 102*, pp. 450-477.

Baker, E., and Stites, R. (1991). Trends in testing in the USA. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing, 1990 yearbook of the politics of educational association* (pp. 139-157). Bristol, PA: Falmer Press.

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., & Shepard, L., (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper# 278. *Economic Policy Institute*.

Barnett, (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children 5*(3), pp. 25-50. Retrieved on December 15, 2013 from http://www.jstor.org/stable/1602366.

Berliner, D., and Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Cambridge, Massachusetts: Perseus Books.

Bieshuvel, S. (1975/ 1999). An examination of Jensen's theory concerning educability, heritability, and population differences. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp.

108-121). New York: Oxford University Press. (Original work published in 1972 in *Psychologia Africana 14* (2).

Block, N. and Dworkin, G. (Eds.). (1976). *The IQ controversy*. New York: Pantheon Books.

Boas, F. (1940). *Race, Language, and Culture*. New York: The Free Press.

Bond, H. (1924a). What the army "intelligence" tests measured. *Opportunity, 2*, 197-202.

Bond, H. (1924b). Intelligence tests and propaganda. *Crisis, 28*(2), 61-64. Retrieved on 8/30/2010 from https://illiad.lib.ku.edu/KKU/illiad.dll?SessionID=C185953238V&Action=10&Form=75&Value=1114673

Bond, H. (1934/ 1966). Chapter XV: Capacity. *The Education of the Negro in the American Social Order*, (pp. 305-330). New York: Octagon Press, Inc.

Bond, H. M. (1956). *A Study of the intelligence of congressmen who signed the southern manifesto as measured by I. Q. tests administered by the Army to them and to their constituents, and by the American Council on Education psychological examinations as administered to, and reportedby, their colleges*. Papers of the National Association for the Advancement of Colored People, Manuscript Division, Library of Congress, Washington, D. C. (This was an unedited manuscript, scanned and sent to me as a PDF).

Boring, (1923). Intelligence as the tests test it. *The New Republic (June),* 35-37

Brace, C. L. (1964). On the race concept. *Current Anthropology, 5*(4), pp. 313-318, 319-320.

Brace, C. L. and Livingstone, F. B. (1971/ 1975/ 1999). On creeping Jensenism. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 207-229). New York: Oxford University Press. (Originally published in 1971 in C. L. Brace, G. R. Gamble, and J. T. Bond (Eds.), *Race and Intelligence*).

Bracey, G. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. *The Phi Delta Kappan 68*(9) pp. 683-686.

Bracey, G. (1996). The sixth Bracey report on the condition of public education. *The Phi Delta Kappan, 78*(2), pp. 127-138. Retrieved on 1/17/2011 from http://www.jstor.org/stable/20405729

Bracey, G. (2001) The 11th Bracey report on the condition of public education. *The Phi Delta Kappan, 83*(2), pp. 157-169. Retrieved on 1/17/2011 from http://www.jstor.org/stable/20440084

Brigham, C. (1923). *A study of American intelligence*. Princeton, New Jersey: Princeton University Press.

Brinkley, A. (2004) *The unfinished nation: A concise history of the American people Volume II: from 1865, (4th ed)*. New York: McGraw Hill Companies, Inc.

Brown v. Board of Educ., 347 U.S. 483 (1954). Retrieved on 9/4/2014 from http://heinonline.org.proxy.library.umkc.edu/HOL/Page?handle=hein.usreports/usrep347&div=45&collection=usreports&set_as_cursor=0&men_tab=srchresults#557.

Bush, G. W. (September 2, 2004). President George W. Bush's Acceptance Speech to the Republican National Convention. *The Washington Post.* Retrieved from http://www.washingtonpost.com/wp-dyn/articles/A57466-2004Sep2.html

Campbell, D. (1975). Assessing the impact of planned social change. In Lyons, C. M. (ed) *Social research and public policies*. Hanover, New Hampshire: Public Affairs Center, Dartmouth College.

Cawelti, G. (2006). The side effects of NCLB. *Education Leadership 64*(3) pp. 64-68.

Clinton, W. J. (2000). Remarks and a question-and-answer session with the Education Writers Association in Atlanta, Georgia. *Weekly Compilation of Presidential Documents*, *36*(15), 819-827.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, S., and York, R. (1966). *Equality of opportunity*. Washington, D.C.: Office of Education, Department of Health, Education, and Welfare.

Commission for Excellence in Education (1983). *A nation at risk. The imperative for educational reform: An open letter to the American people and the secretary of education*. Washington D.C.: U.S. Government Printing Office.

Common Core State Standards Initiative (June 2010). English language arts and literacy in history / social studies, science, and technical subjects. Retrieved on 9/6/2014 from http://www.corestandards.org/ELA-Literacy/.

Coon, C. (1962). *The origin of races*. Oxford, England: Knopf.

Cravens, H. (1990/1987b). Applied science and public policy: The Ohio bureau of juvenile research and the problem of juvenile delinquency. In Michael Sokal, (Ed), *Psychological testing and American society 1890-1930* (pp. 158-194). New York: Rutgers University Press.

Cravens, H. (1988). *The triumph of evolution: The heredity-environment controversy, 1900-1941* (second edition). Baltimore, Maryland: The Johns Hopkins University Press. (Original work published 1978 as *The triumph of evolution: American scientists and the heredity-environment controversy, 1900-1941*)

Cremin, L. (1961/ 1964). *The transformation of the school: Progressivism in American education 1876-1957*. New York: Vintage Books (originally published in 1961).

Darling-Hammond, L. (1990). Achieving our goals: Superficial or structural reforms? *The Phi Delta Kappan International 72*(4), pp. 286-295.

Darling-Hammond, L. (1994). National standards and assessments: Will they improve education? *American Journal of Education 102*(4), pp. 478-510.

Darling-Hammond, L. (2000). New standards and old inequalities: School reform and the education of African American students. *The Journal of Negro Education 69*(4), pp. 263-287.

Darling-Hammond, L. (2004). From "separate but equal" to "no child left behind": The collision of new standards and old inequalities. In D. Meier and G. Wood (Eds.) *Many Children Left Behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston, Massachusetts: Beacon Press/ The Forum for Education and Democracy pp. 3-32

Darling-Hammond, L. (2006). No Child Left Behind and high school reform. *Harvard Educational Review, 76*(4), 642-667.

Darling-Hammond, L. and Berry, B. (1988). *The evolution of teacher policy.* Madison, Wisconsin: The Center for Policy Research in Education of the RAND Corporation.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011). Getting Teacher Evaluation Right: A Background Paper for Policy Makers. *National Academy of Education (NJ1)*.

Darwin, C. (1859/ 1902). *On the origin of species by means of natural selection or the preservation of favored races in the struggle for life*. London: Grant Richards (originally published in 1859).

Davis, A. (1948). *Social-class influences upon learning*. Cambridge, Massachusetts: Harvard University Press.

DeBray, E. H. (2005). Partisanship and ideology in the ESEA reauthorization in the 106th and 107th Congresses: Foundations for the new political landscape of federal education policy. *Review of Research in Education*, 29-50.

Debray-Pellot, E. H. and McGuinn, P. (2009). The new politics of education: Analyzing the federal education policy landscape in the post-NCLB era. *Educational Policy 23*(15) pp. 15-42. Retrieved from http://epx.sagepub.com/cgi/content/abstract/23/1/15.

Dewey, J. (1922/ 2008a). Mediocrity and Individuality. In Boydston, J., and Levine, B. (Eds.) *The middle works of John Dewey, Volume 13, 1899-1924: Journal articles, essays, and miscellany published in the 1921-1922 period.* pp. 289-294. (Originally published in *The New Republic* (33), pp. 35-37).

Dewey, J. (1922/ 2008b). Individuality, Equality and Superiority. In Boydston, J., and Levine, B. (Eds.) *The middle works of John Dewey, Volume 13, 1899-1924: Journal articles, essays, and miscellany published in the 1921-1922 period.* pp. 295-300. (Originally published in *The New Republic* 33, pp. 61-33.

Dewey, J. (1938/ 1997). *Experience and Education*, New York: Touchstone. (Original work published 1938).

Dobzhansky, T. (1973). *Genetic diversity and human equality*. New York: Basic Books.

Dunbar, S., Koretz, D., and Hoover, H. (1991). Quality control in the use of performance assessment. *Applied Measurement in Education, 4*, pp. 289-303.

Duncan, A. (July 29, 2013). Statement from U.S. Secretary of Education Secretary Duncan on House ESEA Reauthorization Bill H.R. 5 [press release]. Retrieved on May 24, 2014 from https://www.ed.gov/news/press-releases/statement-us-secretary-education-secretary-duncan-house-esea-reauthorization-bil.

Duncan, O. D. (1968). Socioeconomic Background and Occupational Achievement: Extensions of a Basic Model. Final Report.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. (1951). *Intelligence and cultural differences: A study of cultural learning and problem-solving*. University of Chicago Press.

Elementary and Secondary Education Act, 20 U.S.C § 6301 (1965).

Evers, W. (2001) Chapter 9: Standards and accountability. In T. Moe (Ed.), *A primer on America's schools*, Stanford, California: Hoover Institute Press Publications pp. 205-247.

Fass, P. (1980). The IQ: A cultural and historical framework. *American Journal of Education 88*(4), pp. 431-458.

Finn, C. and Ravitch, D. (1996). Part IV: Instruction: The tyranny of dogma. *A report from the educational excellence network to its educational policy committee and the American people. Thomas B. Fordham Institute.*

Finn, C. E. (2002). Real accountability in K-12 education. *School accountability*, 23-46.

Finn, C. E., Kanstoroom, M., Rothstein, R., & Honig, B. (2001). State academic standards. *Brookings papers on education policy*, 131-179.

Finn, C. E., Manno, B. V., & Vanourek, G. (2001). The radicalization of school reform. *Society*, *38*(4), 58-63.

Fredrickson, G. (1981). *White supremacy: A comparative study in American and South African history.* London: Oxford University Press.

Freire, P. (1990). *Pedagogy of the oppressed*. New York: Continuum.

Fried, M. (1968). The need to end the pseudoscientific investigation of race. In Mead, M., et al., (eds.) *Science and the concept of race*. New York: Columbia University Press.

Fuhrman, S., and Malen, B. (Eds.). (1991). *The politics of curriculum and testing: The 1990 yearbook of the politics of education association.* Bristol, Pennsylvania: Falmer Press, Ltd.

Fusarelli, L. (2004). The potential impact of the No Child Left Behind Act on Equity and Diversity in American Education. *Educational Policy 18*(1), 71-94.

Gaddis, J. (2002). *The landscape of history: How historians map the past*. Oxford, New York: Oxford University Press.

Gage, N. (1972). I.Q. heritability, race differences, and educational research. *Phi Delta Kappan*, pp. 297-307.

Galton, F. (1870). *Hereditary Genius: An inquiry into its laws and consequences*. New York: D. Appleton and Co.

Gamoran, A. (2007b). Chapter 1: Introduction: Can standards-based reform help reduce the poverty gap in education? In A. Gamoran (Ed.) *Standards-based reform and the poverty gap: Lessons for No Child Left Behind.* Washington D.C.: Brookings Institution Press, pp. 3-16.

Gamoran, A. (Ed.). (2007). *Standards-based reform and the poverty gap: Lessons for No Child Left Behind.* InWashington D.C.: Brookings Institution Press.

Garn, S. (1964). *Culture and the direction of human evolution*. Detroit, Michigan: Wayne State University Press.

Garrison, J. (2012). Individuality, equality, and creative democracy—the task before us. *Journal of Education 118*(3) 369-379. Retrieved on 1/28/2014 from http://www.jstor.org/stable/10.1086/664739.

Gerstner, L., Semerad, R., Doyle, D., and Johnston, W. (1994). *Reinventing education: Entrepreneurship in America's public schools*. New York: Dutton.

Gladwell, M. (2008). *Outliers: The story of success.* New York: Malcolm Gladdwell and Little Brown and Company.

Goals 2000: Educate America Act Pub. L. No.103-227 (1994), retrieved on August 16, 2011 from http://www2.ed.gov/legislation/GOALS2000/TheAct/index.html.

Goddard, H. (1922). *Human efficiency and levels of intelligence: Lectures delivered at Princeton University on April 7, 8, 10, and 11, 1910*. Princeton, New Jersey: Princeton University Press.

Gordon, E., and Green, D. (1975/ 1999). An affluent society's excuses for inequality: Developmental, economic, and educational. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp.122-152). New York: Oxford University Press. (Original work published in 1974 in *American Journal of Orthopsychiatry 44*(1) 4-18).

Gottfredson, L. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence 24*(3), 13-23. (Originally published on December 13, 1994 in *The Wall Street Journal)*.

Gould, S. (1974/ 1975/ 1999). Racist Arguments and IQ. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 184-189). New York: Oxford University Press. (Originally published in 1974 in *Natural History Magazine*).

Gould, S. (1981/ 1996). *The mismeasure of man*. New York: W. W. Norton & Company.

Grant, M. (1916). *The passing of the great race or the racial basis of European history.* New York: Charles Scribner's Sons.

Gregg, T., and Sanday, P. (1971). Genetic and environmental components of differential intelligence. In Brace, C. L., Gamble, G., and Bond, J. (Eds.) *Race and intelligence: Anthropological studies*, (8). Washington, D. C.: American Anthropological Association.

Gresson III, A. (1996/ 1997). Prelude. Kincheloe, J., Steinberg, S., and Gresson, A. (Eds.). (1996/ 1997). *Measured lies: The Bell Curve examined*. New York: St. Martin's Press, pp. ix-x.

Haas, E., Wilson, G., Cobb, C., & Rallis, S. (2005). One hundred percent proficiency: A mission impossible. *Equity & excellence in education*, *38*(3), 180-189.

Haertel, E. and Calfee, R. (1983). School achievement: Thinking about what to test. *Journal of Education Measurement, 20*, pp. 119-132.

Haertel, E. and Herman, J. (2005a). A historical perspective on validity arguments for accountability testing. CSE Report 654. *Los Angeles California: Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.* Retrieved on 4/10/2009 from http://www.eric.ed.gov/PDFS/ED488709.pdf.

Hammond, B. (May 12, 2014). Oregon teachers union calls for moratorium on Common Core reading, math test. *The Oregonian*. Retrieved on 9/6/2014 from http://www.oregonlive.com/education/index.ssf/2014/05/oregon_teachers_union_calls_fo.html.

Hanushek, E. (1989). Expenditures, efficiency, and equity in education: The federal government's role. *The American Economic Review 79*(2) 46-51. Retrieved on 1/1/2013 from http://www.jstor.org/stable/1827728.

Hanushek, E. (1994). Making America's schools work: This time money is not the answer. *The Brookings Review 12*(4) 10-13.

Hanushek, E. (2003). The failure of input-based schooling policies. *The Economic Journal 113*(485) F64-F98. Retrieved on 1/1/2013 from http://www.jstor.org/stable/3590139.

Hanushek, E., and Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), pp. 297-327. Retrieved on 8/7/2010 from http://www.jstor.org/stable/3326211.

Harvey, D. (2005). *A brief history of neoliberalism.* New York: Oxford University Press.

Hazlett, J. (1974). *A history of the National Assessment of Educational Progress, 1963-1973*. Ed. D. Dissertation, University of Kansas.

Heckman, J. (1995). Lessons from the bell curve. *The Journal of Political Economy 103*(5) 1091-1120.

Heckman, J., & Kautz, T. (2013*). Fostering and measuring skills: Interventions that improve character and cognition* (no. w19656). National Bureau of Economic Research.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review, 103*(6), pp. 2052-2086.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Herrnstein, R., & Murray, C. (1994) *The bell curve: Intelligence and class structure in American life.* New York: A Free Press Book.

Heubert, J. P., & Hauser, R. M. (1999). High stakes: Testing for tracking, promotion, and graduation. *Washington, DC: National Academy Press.*

Honzik, M. (1957). Developmental studies of parent-child resemblance in intelligence. *Child Development, 28*, pp. 215-228.

Huxley, J. S. (1938). Clines: an auxiliary taxonomic principle. *Nature,142*(3587), 219-220.

Imber, M., and van Geel, T. (2004). *Education Law, 3rd ed.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Improving America's Schools Act, Pub. L. No.103-382 (1994).

Jackson, J., and Weidman, N. (2006). *Race, Racism, and Science: Social Impact and Interaction*. New Brunswick, New Jersey: Rutgers University Press. (Volume in the Science and Society Series, M. Largent (Series Ed.).

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Review, 39*, (1), pp. 1-123.

Kagan, J. (1971/ 1975/ 1999). The magical aura of the IQ. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 101-107). New York: Oxford University Press. (Originally published in 1971 in *The Saturday Review* pp. 92-93).

Kamin, L. (1999). Behind the Curve. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 397-407). New York: Oxford University Press.

Kanstoroom, M., and Finn Jr., C. (Eds.). (1999). *Better teachers, better schools*. Washington, D.C.: Thomas B. Fordham Foundation.

Karier, C. (1986). *The individual, society, and education: A history of American educational ideas* (2nd ed.). Urbana and Chicago, Illinois: University of Illinois Press.

Karp, S. (2004). NCLB's selective vision of equality: Some gaps count more than others. In D. Meier and G. Wood (Eds.) *Many Children Left Behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston, Massachusetts: Beacon Press/ The Forum for Education and Democracy pp. 53-65

Kearns, D. (1988). An education recovery plan for America. *Phi Delta Kappan 69*(8), 565-570.

Kearns, D. and Doyle, D. (1989). *Winning the brain race: A bold plan to make our schools competitive.* San Francisco, CA: Institute for Contemporary Studies.

Kett, J. (2013). *Merit: The history of a founding ideal from the American Revolution to the twenty-first century*. Ithaca, New York: Cornell University Press.

Kincheloe, J., Steinberg, S., and Gresson, A. (Eds.). (1996/ 1997). *Measured lies: The Bell Curve examined*. New York: St. Martin's Press.

Klineberg, O. (1928). *An experimental study of speed and other factors in" racial" differences* (No. 93). University Microfilms International.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education*, *6*(1), 18-42.

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

Kohn, A. (2004). NCLB and the effort to privatize public education. In D. Meier and G. Wood (Eds.) *Many Children Left Behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston, Massachusetts: Beacon Press/ The Forum for Education and Democracy pp. 79-100

Kohn, A. (2004a). NCLB and the effort to privatize public education. In D. Meier and G. Wood (Eds.) *Many Children Left Behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston, Massachusetts: Beacon Press/ The Forum for Education and Democracy pp. 79-100

Kohn, A. (2004b). Test today, privatize tomorrow: Using accountability to 'reform' public schools to death. *The Phi Delta Kappan, 85*(8, pp. 568-577. Retrieved on 1/17/2011 from http://www.jstor.org/stable/20441647

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, Massachusetts: Harvard University Press.

Kozol, J. (1991). *Savage inequalities: Children in America's schools*. New York: Crown Publishers, Inc.

Lane, C. (1999). The tainted sources of the bell curve. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 408-424). New York: Oxford University Press.

Layton, L. (June 10, 2014). Gates Foundation urges delay in using tests for teacher evaluation. *Washington Post*. Retrieved on 9/6/2014 from http://www.washingtonpost.com/local/education/gates-foundation-urges-delay-in-using-tests-for-teacher-evaluation/2014/06/10/d037c7fa-f0e1-11e3-914c-1fbd0614e2d4_story.html.

Lee, V. E., Brooks-Gunn, J., Schnur, E., & Liaw, F. R. (1990). Are Head Start Effects Sustained? A Longitudinal Follow-up Comparison of Disadvantaged Children Attending Head Start, No Preschool, and Other Preschool Programs. *Child development*, *61*(2), 495-507.

Lemann, N. (2000). *The big test: The secret history of the American Meritocracy*. New York: Farber, Strauss and Giroux.

Lewis, A. C. (2002). A horse called NCLB. *Phi Delta Kappan, 84*(3), 179-180.

Lewontin, R. (1970/ 1975/ 1999). Race and intelligence. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 230-247). New York: Oxford University Press. (Originally published in 1970 in *Science and Public Affairs*, 2-8).

Lieberman, L., Littlefield, A., and Reynolds, L. (1999). The debate over race: Thirty years and two centuries later. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 46-90). New York: Oxford University Press.

Lippmann, W. (1922a). The mental age of Americans. *New Republic 32*(412), 213-215.

Lippmann, W. (1922b). The mystery of the "A" men. *New Republic 32*(413), 246-248.

Lippmann, W. (1922c). The reliability of intelligence tests. *New Republic 32*(414), 275-277.

Lippmann, W. (1922d). The Abuse of the Tests. *New Republic 32*(415), pp. 297-298.

Lippmann, W. (1922e). Tests of Hereditary Intelligence. The New Republic 33(416) pp, 328-330

Lippmann, W. (1922f). The future for tests. New Republic 33(417), 9-11.

Lippmann, W. (1923a). The Great Confusion. The New Republic 34 (January 3, 1923), pp. 145-146.

Loveless, T. (2007). The peculiar politics of No Child Left Behind. In A. Gamoran (Ed.), *Standards based reform and the poverty gap: Lessons for No Child Left Behind.* Washington, D.C.: Brookings Institution Press, pp.253-285.

Luria, S. E. (1974/ 1975/ 1999). What can biologists solve? In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 91-100). New York: Oxford University Press. (Originally published in 1974 in *The New York Review of Books*, 27-28).

Lynn, R. (1991). Race differences in intelligence: A global perspective. *Mankind Quarterly 31*, pp. 254-296.

Madaus, G. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46. Retrieved on 3/30/2010 from http://www.jstor.org/stable/1492818.

Madaus, G., Russell, M., and Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society.* Charlotte, North Carolina: Information Age Publishing, Inc.

Marx, K. (1843/ 1970). *Critique of Hegel's "Philosophy of the Right."* Translated by Joseph O'Malley. New York: Oxford University Press.

McGuinn, P. (2005). The national schoolmarm: "No child left behind" and the new educational federalism. *Publius: The state of American federalism. 35*(1).

McNeil, L. (1988a). Contradictions of control, Part 1: Administrators and teachers. *The Phi Delta Kappan, 69*(5), 333-339. Retrieved on 3/30/2010 from http://www.jstor.org/stable/20403627

McNeil, L. (1988b). Contradictions of control, Part 2: Teachers, students, and curriculum. *Phi Delta Kappan, 69*(6), 432-438. Retrieved on 3/30/2010 from http://www.jstor.org/stable/20403666

McNeil, L. (1990). Reclaiming a voice: American curriculum scholars and the politics of what is taught in schools. *Phi Delta Kappan, 71*(7), 517-519. Retrieved on 3/30/2010 from http://www.jstor.org/stable/20404199.

McNeil, L. (2000). Creating new inequalities: Contradictions of reform. *Phi Delta Kappan, 81*(10), 728-734. Retrieved on 3/30/2010 from http://www.jstor.org/stable/20439779.

Mead, G. H. (1964). On social psychology. Chicago, Illinois: University of Chicago Press.

Meier, D. (2004). NCLB and democracy. In D. Meier and G. Wood (Eds.) *Many Children Left Behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston, Massachusetts: Beacon Press/ The Forum for Education and Democracy pp. 66-78.

Meier, D., and Wood, G. (Eds.). (2004). *Many children left behind: How the No Child Left Behind is damaging our children and our schools*. Boston Massachusetts: Beacon Press/The Forum for Education and Democracy.

Merriam, S. (2009). *Qualitative research: A guide to design and implementation*. San Francisco, California: Jossey-Bass.

Mincer, J. (1972). *Schooling, experience, and earnings*. New York: Columbia University Press

Minton, H. (1987/ 1990). Lewis M. Terman and mental testing: In search of the democratic ideal. In Michael Sokal, (Ed), *Psychological testing and American society 1890-1930* (pp. 95-112). New York: Rutgers University Press.

Montagu, A. (1941). The concept of race in the human species in light of genetics. *Journal of Heredity, 23*. Pp. 243-247.

Montagu, A. (1945). Intelligence of Northern Negroes and Southern Whites in the First World War. *The American Journal of Psychology, 58*(2), 161-188. Retrieved on 8/29/2010 from http://www.jstor.org/stable/1417844.

Montagu, A. (1975/ 1999a). *Introduction*. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 1-18). New York: Oxford University Press.

Montagu, A. (1975/ 1999b). The IQ mythology. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 29-45). New York: Oxford University Press.

Montagu, A. (1975/ 1999c). Intelligence, IQ, and race. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 190-206). New York: Oxford University Press.

Montagu, A. (1997). *Man's most dangerous myth, 6th ed.* Walnut Creek, CA: Alta Mira Press.

Montagu, A. (Ed.). (1975/ 1999). *Race and IQ: Expanded edition*. Oxford, New York: Oxford University Press.

Murdoch, S. (2007). *IQ: A smart history of a failed idea.* Hoboken, New Jersey: John Wiley and Sons, Inc.

National Education Association Commission on the Reorganization of Secondary Education. (1918). *Cardinal principles of secondary education: A report of the commission on the reorganization of secondary education, appointed by the National Education Association.* Washington, D.C.: Government Printing Office.

National Research Council, Committee on Appropriate Test Use, Board on Testing and Assessment (1999). *High stakes: Testing for tracking, promotion, and graduation.* Heubert, Jay P., and Robert M. Hauser, (Eds.). Washington, D.C.: National Academy Press.

Neal, D., and Johnson, W. (1994). *The role of pre-market factors in Black-white wage differences*. Chicago, Illinois: University of Chicago.

Neustadt, R., and May, E. (1986). *Thinking in time: The uses of history for decision makers*. New York: The Free Press.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Ohanian, S. (1999). *One size fits few: The folly of educational standards*. Portsmouth, New Hampshire: Heinemann.

Ohanian, S. (2000). Goals 2000: What's in a name? *Phi Delta Kappan, 81*(5), 344-355. Retrieved on 3/30/2010 from http://www.jstor.org/stable/20439663.

Ohanian, S. (2003). Capitalism, calculus, and conscience. *Phi Delta Kappan, 84*(10), 736-747. Retrieved on 6/22/2010 from http://www.jstor.org/stable/20440474.

Peterson, P. E., & Hess, F. M. (2008). Few states set world-class standards. *Education Next, 8*(3), 70-73.

Phillips, D. A., & Shonkoff, J. P. (Eds.). (2000). *From Neurons to Neighborhoods: The Science of Early Childhood Development*. National Academies Press.

PL 103- 382 *Improving America's Schools Act of 1994*.

Raven, I., Summers, B., Birchfield, M., Brosier, G., Burciaga, L., Bykrit, B., et al. (1990). *Manual for Raven's Progressive Matrices and Vocabulary Scales - Research supplement: no. 3. American and international norms (2nd ed.)*. Oxford, England: Oxford Psychologists Press.

Ravitch, D (1993). Launching a revolution in standards and assessments. *Phi Delta Kappa International 74*(10), pp. 767-772.

Ravitch, D. (1994). Somebody's children: Expanding educational opportunities for all America's children. *The Brookings Review*, 4-9.

Ravitch, D. (1996). The case for national standards and assessments. *The Clearing House, 69*(3), 134-135.

Ravitch, D. (1999). Student performance: The national agenda in education. *The Brookings Review*, 12-16.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Ravitch, D. (2011, May 31). Waiting for a school miracle. *The New York Times*. Retrieved on 5/24/2014 from http://www.nytimes.com/2011/06/01/opinion/01ravitch.html?_r=0. (A print version appeared on June 1, 2011 on p. A27 of the *New York Times*).

Ravitch, D. (June 30, 2012). Are standardized tests worthless? *Diane Ravitch's Blog: A site to discuss better education for all*. [Blog post]. Retrived on 9/6/2014 from http://dianeravitch.net/2012/07/30/are-standardized-tests-worthless/.

Ravitch, D. (April 6, 2014). Testing in Texas: The true story of a professor, angry moms, and legislators who listened. *Diane Ravitch's Blog: A site to discuss better education for all*. [Blog post]. Retrived on 9-6-2014 from http://dianeravitch.net/2014/04/06/testing-in-texas-the-true-story-of-a-professor-angry-moms-and-legislato-rs-wholistened/.

Ravitch, D. (September 5, 2014a). Peter Greene deconstructs think tanker's ideas on teacher evaluation. *Diane Ravitch's Blog: A site to discuss better education for all*. Retrieved on 9/6/2014 from http://dianeravitch.net/2014/09/05/peter-greene-deconstructs-think-tankers-ideas-about-teacher-evaluation/.

Ravaitch, D., (September 5, 2014b). What happened to the scholar who challenged Pearson? *Diane Ravitch's Blog: A site to discuss better education for all*. Retrived on 9-6-2014 from http://dianeravitch.net/2014/09/05/what-happened-to-the-scholar-who-challenged-pearson/.

Reed, J. (1990/1987). Robert M. Yerkes and the mental testing movement. In Michael Sokal, (Ed.), *Psychological testing and American society 1890-1930* (pp. 75-94). New York: Rutgers University Press.

Reid, K. S. (2005). Civil Rights Groups Split over NCLB: Accountability Provisions Stirring Heated Debate. *Education Week*, *25*(1), 1-20.

Rothstein, R. (2004). *Class and schools: Using social, economic and educational reform to close the Black-White achievement gap.* Washington, D.C.: Economic Policy Institute

Resmovits, (April 30, 2013). Common Core stakes moratorium proposed by union as national standards face backlash. *Huffington Post*. Retrieved on 9/6/2014 from http://www.huffingtonpost.com/2013/04/30/common-core-moratorium-teacher-evaluations_n_3187419.html.

Rothstein, R. (2008). Leaving "No Child Left Behind" Behind. *American Prospect*, *19*(1), 50.

Rothstein, R., Jabobson, R., and Wilder, T. (2008). Grading Education: Getting Accountability Right. New York, New York and Washington, DC: Economic Policy Institute/ Teachers College Press.

Rury, J. (1988). Race, region, and education: An analysis of Black and White scores on the 1917 army alpha intelligence test. *The Journal of Negro Education 57*(1), pp. 51-65.

Rury, J. (1995) IQ Redux. *History of Education Quarterly 35*(4), 423-438. Retrieved on September 11, 2010 from http://www.jstor.org.www2.lib.ku.edu:2048/stable/pdfplus/369579.pdf

Rury, J. L. (2012). Two Cheers for NCLB, and Questions for Professor Garrison. *American Journal of Education*, *118*(3), 385-388.

Rury, J., & Akaba, S. (2014 in print). The Geo-spatial distribution of educational attainment: Cultural capital and uneven development in metropolitan Kansas City, 1960-1980. *HISTOIRE & MESURE* (*History and Measurement*) XXIX-1, pp. 219-246.

Ryan, A. (1999). Bad science, worse politics. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 379-396). New York: Oxford University Press.

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it.* Cambridge, Massachusetts: Perseus Publishing.

Salzman, A. (2006, February 1). NAACP is Bush Ally in School Suit Versus State. *The New York Times*, B3(L). Retrieved on 1/28/2014 from *Academic One File* at http://go.galegroup.com/ps/i.do?id=GALE%7CA141516176&v=2.1&u=ksstate_ukans&it=r&p=AONE&sw=w&asid=54dbe612ead50a5d64c05b49ddab34a8.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences13*, pp. 279-280.

Samelson, F. (1990/1987). Was early mental testing (a) racist inspired, (b) objective science, (c) a technology for democracy, (d) the origin of multiple-choice exams, (e) none of the above? (mark the right answer). In M. Sokal, (Ed), *Psychological testing and American society 1890-1930* (pp. 113-127). New York: Rutgers University Press.

Sanday, P. (1972/ 1975/ 1999). On the causes of IQ differences between groups and implications for social policy. In A. Montagu (Ed.), *Race and IQ, expanded edition* (pp. 276-307). New York: Oxford University Press. (Originally published in 1972 in *Human Organization 31*(4), 411-424.

Shavelson, R., Baxter, G., and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, pp. 215-232.

Skodak, M., and Skeels, H. (1949). A final follow-up study of one hundred adopted children. *Journal of Genetic Psychology, 75*, pp. 85-125.

Smith, M., and O'Day, J. (1991). Systemic school reform. In S. Fuhrman and B. Malen (Eds.). *The politics of curriculum and testing*, pp. 233-267. London: Falmer Press.

Sokal, M. (1990/ 1987). Introduction: Psychological testing and historical scholarship—questions, contrasts, and context. In M. Sokal, (Ed), *Psychological testing and American society 1890-1930,* pp. 1-20). New York: Rutgers University Press.

Sokal, M. (1990/ 1987). James McKeen Cattell and mental anthropometry: Nineteenth-century science and reform and the origins of psychological testing. In M. Sokal, (Ed), *Psychological testing and American society 1890-1930* (pp. 21-45). New York: Rutgers University Press.

Sokal, M. (Ed.). (1990/ 1987). Psychological testing and American society 1890-1930. New York: Rutgers University Press.

Southern Poverty Law Center Web site, "The Pioneer Fund" Retrieved on December 7, 2013 from http://www.splcenter.org/get-informed/intelligence-files/groups/pioneer-fund

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201-292.

Spencer, H. (1864). *The principles of biology*. London: Williams and Northgate.

Strauss, V. (November 7, 2013). The biggest weakness of the Common Core Standards. *The Washington Post* [Blog Post]. Retrieved on 9/5/2014 from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/11/07/the-biggest-weakness-of-the-common-core-standards/

Terman, L. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Cambridge, Massachusetts: The Riverside Press.

Terman, L. (1919). *The intelligence of school children: how children differ in ability the use of mental tests in school grading and the proper education of exceptional children*. Cambridge, Massachusetts: The Riverside Press.

Terman, L. (1920). The use of intelligence tests in the grading of school children. Journal of Educational Research 1(1), pp. 20-32.

Terman, L. (1922b). The great conspiracy, or the impulse imperious of intelligence testers, psychoanalyzed and exposed by Mr. Lippmann. *New Republic 33*(December 27, 1922), 116-120.

Terman, L. (January 17, 1923) Letter to *The New Republic*.

Terman, L. (June 1922a). The psychological Determinist: or Democracy and the I.Q. *The Journal of Educational Research*, 6(1), pp. 57-62.

The Improving America's Schools Act (1994). Amendments to the 1965 Elementary and Secondary Education Act 20. U.S.C. 2701 H.R. 6

Thomas, W. (1982). Black intellectuals' critique of early mental testing: A little-known saga of the 1920s. *American Journal of Education, 90*(3), 258-292.

Thorndike, E., (1908). *The elimination of pupils from school. Bulletin No. 4, whole number*. Washington, D. C.: Government Printing Office

Thorndike, E., (1919). Scientific personnel work in the army. *Science, New Series 49*(1255), pp. 53-61.

Tilly, C. (1998). *Durable inequality*. Berkley, CA and Las Angeles, CA: University of California Press.

Title I Evaluation and Reporting System (TIERS), 1974

Toch, T. (2011). Who rules? The Wilson Quarterly 35(4), pp. 43-47.

Truth in American Education. (March, 2011). *Common Core $tate $tandards: What parents, taxpayers and school boards should know, that perhaps they aren't being told*. [Online Newsletter] Retrieved on 9-5-2014 from

Tyack, D. (1974). *The one best system: A history of American urban education*. Harvard University Press.

Tyack, D., and Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform.* Cambridge, Massachusetts: Harvard University Press.

UNESCO. (1997). *Statistical yearbook 1997*. Paris: UNESCO.

Ujifusa, A. (June 6, 2014). Days apart, two states opt to replace Common Core. *Education Week, 33*(35), pp. 24, 26.

Von Mayrhauser, R. T.. (1987/ 1990). The manager, the medic, and the mediator: The clash of professional psychological styles and the wartime origins of group mental testing. In Michael Sokal, (Ed), *Psychological testing and American society 1890-1930* (pp. 128-157). New York: Rutgers University Press.

Walberg, H. J. (1994). Educational productivity: Urgent needs and new remedies. *Theory into Practice, 33*(2), 75-82. Retrieved on 1/19/2013 from http://www.jstor.org/stable/1476566

Walberg, H. J., (2001). Chapter 3: Achievement in American schools. . In T. Moe (Ed.), *A primer on America's schools*, Stanford, California: Hoover Institute Press Publications pp. 43-67.

Walberg, H. J., Bishop, J., & Hannaway, J. (1998). Uncompetitive American schools: Causes and cures. *Brookings papers on education policy*, 173-226. Retrieved on 1/19/2013 from http://www.jstor.org/stable/20067197.

Wallace (1968). Schools in revolutionary and conservative societies. In E. M. Lloyd-Jones and N. Rosenau (Eds.), *Social and cultural foundations of guidance: A sourcebook,* pp 193-203. New York: Holt, Rinehart and Winston, Inc.

Weber, M. (1905/ 1958). *The Protestant ethic and the spirit of capitalism*. Translated by Talcott Parsons. New York: Scribner.

Whipple, G. M. (1922). Educational determinism: A discussion of Professor Bagley's address at Chicago. *School and Society 15*(388), 600-602.

Wiebe, R. (1967). *The search for order: 1877-1920*. New York: Hill and Wang.

Wiggins, G. (1992). Creating tests worth taking. *Education Leadership, 49*(8), pp. 26-33.

Wood, G. (2004). A view from the field: NCLB's effects on classrooms and schools. In D. Meier and G. Wood (Eds.) *Many Children Left Behind: How the No Child Left Behind Act is damaging our children and our schools.* Boston, Massachusetts: Beacon Press/ The Forum for Education and Democracy, pp. 33-52.

Yerkes, R., Haggerty, M., Terman, L. Thorndike, E., and Whipple, G. (1920). National intelligence tests. *The Elementary School Journal 21*(3), p. 239.

Yerkes, R., (1921). *Psychological Testing in the United States Army.* Retrieved on 9/4/2010 from http://books.google.com/books?hl=en&lr=&id=YIH7DBtwZOkC&oi=fnd&pg=PR7&dq=Yerkes+results+of+psychological+examination&ots=QwtPLGDQRp&sig=xmZL-bKGBHKIP2PjMjJl5xTTSco#v=onepage&q=Yerkes%20results%20of%20psychological%20examination&f=false

Yoakum, C., and Yerkes, R. (Eds.). (1920). *Army Mental Tests*. New York: Henry Holt and Company. Retrieved on 9/4/2010 from http://www.archive.org/details/armymentaltests002887mbp

Young, K. (1923). *Mental differences in certain immigrant groups: Psychological tests of South Europeans in typical California schools with bearing on the educational policy and on the problems of racial contacts in this country.* Eugene, Oregon: The University, University Press.

Zenderland, L. (1990/1987). The debate over diagnosis: Henry Herbert Goddard and the medical acceptance of intelligence testing. In Michael Sokal, (Ed), *Psychological testing and American society 1890-1930* (pp. 46-74). New York: Rutgers University Press.