THE USE OF STANDARDIZED TEST SCORES: AN HISTORICAL PERSPECTIVE

by

KARLA LAIRSEY SWAFFORD

(Under the Direction of William Wraga)

ABSTRACT

The field of educational tests and measurements is a broad one, complete with a variety of perspectives about how educators should approach assessment of student ability and learning. The standardized testing movement is used by politicians and policymakers to hold schools accountable. Each decade of the twentieth century and the first decade of the twenty-first century have seemingly brought a different argument as to what the shortcomings of American schools are and what to do to solve that problem. However, there has been one constant denominator: assessing students' learning through a standardized test. Recently legislation in the United States, including the national No Child Left Behind Act and Georgia's A+ Reform Act, has begun to dictate that students, teachers, administrators, and schools as a whole be judged on the basis of a single test score. This study uses historical research methods to explore the history of standardized testing in the United States and what lead test developers have held to be the appropriate and inappropriate uses of test scores. Additionally, this study provides background on seven standardized tests and their appropriate uses. The seven assessments include the California Achievement Test, the Comprehensive Test of Basic Skills, the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the National Assessment for Educational Progress, the Scholastic Aptitude Test, and the Stanford Achievement Test. The study finds that

test scores are properly used for determining student achievement and the effectiveness of an educational program as long as the scores are used along with other pieces of information, such as teacher observations and performance assessment data, to make high-stakes decisions about students. To use a single score to make important decisions such as student promotion, retention, and graduation or about the quality of schools is deemed an inappropriate use by testing experts and test publishers. These findings are in direct opposition to the current policy such as the No Child Left Behind Act of 2001 that requires that schools and students be judged based on one single test score. Such public policy and legislation should be reviewed for their suitability in school reform efforts.

INDEX WORDS: Educational Measurement; Educational Assessment; California Achievement Test; Comprehensive Test of Basic Skills; Stanford Achievement Test; Iowa Test of Basic Skills; Metropolitan Achievement Test; Scholastic Aptitude Test; National Assessment of Educational Progress; No Child Left Behind

THE USE OF STANDARDIZED TEST SCORES: AN HISTORICAL PERSPECTIVE

by

KARLA LAIRSEY SWAFFORD

Bachelor of Arts, Georgia State University, 1993

Master of Arts, University of Georgia, 1999

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF EDUCATION

ATHENS, GEORGIA

2007

THE USE OF STANDARDIZED TEST SCORES: AN HISTORICAL PERSPECTIVE

by

KARLA LAIRSEY SWAFFORD

Major Professor:	Dr. William Wraga
Committee:	Dr. Catherine Sielke Dr. John Dayton

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December, 2007

DEDICATION

This work is dedicated to my family and to God. My family has been unfailing in their support of me, and they have never wavered in their certainty that I could do this. My husband, Roger; my son, Jackson; and my parents, Ralph and Nadean, were a source of constant encouragement to me. God has blessed my life far beyond what I deserve, and He has given me the opportunities and the ability to accomplish much. For this, I am unspeakably grateful.

ACKNOWLEDGEMENTS

I would like to acknowledge my major professor, Dr. William Wraga. He has been a wonderful mentor, never pushing me more than I needed but always there to nudge me along. His influence over my academic career is great, and I am deeply grateful for his support. He has regarded me not only as a doctoral student but also as a woman with a family who always took priority in my life and as a fellow educator with responsibilities that often pulled me away from my research. However, he continually encouraged me and was always ready with a supportive word to keep me plodding along.

I would also like to recognize Dr. John Dayton and Dr. Catherine Sielke for serving on my committee and for their contributions to my work.

Finally, I would like to acknowledge the following individuals from the testing companies and agencies who were invaluable in my gathering of the research represented by this work.

Educational Testing Services—Andres Pollack—Andres spent countless hours pulling from the ETS Archives for me. His assistance was invaluable to me.

The College Board—Wayne Camara

Riverside Publishing—Hala Istanbouli

National Center for Educational Statistics—Arnold Goldstein

TABLE OF CONTENTS

Page

Chapter 1

STANDARDIZED TESTING

Background of the Problem

The field of educational tests and measurements is a broad one, complete with a variety of perspectives and philosophies about how educators should approach assessment of student ability and learning. Testing members of society goes back to 2200 B. C. when the emperor of China began using civil service examinations to determine the fitness of public officials (Wardrop, 1976). This movement spread from China to Europe in the 1790's and finally to the United States in 1883 when the United States Civil Service Commission came into being (Dubois, 1970; Wardrop, 1976). Exams in universities became commonplace in the determination of academic achievement for students in 1219 at the University of Bologna (Wardrop, 1976). By 1540, the Jesuits were considered "pioneers in the systematic use of written tests" with very strict rules of examination (Dubois, 1970, pp. 8-9). These tests were used as a method of placing students and then as a way to evaluate them at the conclusion of a period of instruction (Wardrop, 1976).

Psychological testing that aims to assess not achievement but the abilities of students began during the nineteenth century with Alfred Binet in France, Sir Frances Galton in England, and with James McKeen Cattell in the United States (Wardrop, 1976). These men began the trend of looking at what people had as innate abilities rather than attempting to look at how to measure what students had been taught. To be more precise, they sought to measure the level of intelligence with which a person had been born as a way to sort students rather than merely

assessing how well they had learned the curriculum in which they had been instructed. Thus the rise in the use of "IQ tests" to sort American children commenced. According to Patrick Lemann (2000), "True believers in IQ tests thought they should be given to all American school children, so that the high scorers could be plucked out and given the best schooling and the average and low scorers consigned to a briefer, more limited education" (p. 23). The students whose IQ's were higher were given the most academic and more scholarly education while the other students' education focused mainly on work skills and more practical applications for use in everyday life. Lemann continued, "Americans did not publicly announce that they were sorting their children on the basis of IQ tests" (p. 119), but that is exactly what was happening all over the country.

However, it was the widespread use of the multiple choice test item that gave educational testing its big boost. According to J. L. Wardrop (1976), there were three main reasons for the growth of the industry: new statistical procedures for analyzing and improving tests, faster ways of scoring the tests and reporting the results, and the "institutionalization of testing in American society" (p. 14). In 1922, S. L. Pressey and L. C. Pressey maintained, "It is quite evident that the school people of the country are becoming more and more interested in 'tests'" (p. 20). How prophetic that statement was. From the 1930s forward, educational testing in the United States became the dominant method of determining how much and how well children were learning in American schools. During the 1930s, as textbook publishers like Harcourt Brace Jovanovich and Houghton Mifflin became test publishers (Wardrop, 1976), they also became ever-increasingly in control of the curriculum taught in schools in the United States.

School reform efforts were motivated to use these standardized tests as a way to measure student progress. From the 1950s to the current day, schools have come under fire for not

preparing students to compete internationally. Each decade has seemingly brought a different argument as to what the shortcomings of American schools are and what to do to solve that problem. However, there has been one constant denominator: assessing students' learning through a standardized test. By the time that *A Nation At Risk* was published in 1983, the testing industry already had a foothold in the economy of educating America's youth. But reports like *A Nation At Risk*, which calls for the systematic testing of students, have given the testing industry momentum which is likely unstoppable.

The results of standardized tests are seen as the ultimate way to measure a school's and a teacher's effectiveness. According to Gerald Bracey (1998), standardized testing has a variety of uses and misuses. Monitoring, or measuring student progress, and principal and school board accountability are just two of the uses that Bracey mentions. However, three blatant misuses of test scores are also cited by Bracey: to diagnose what knowledge and skills students may lack, to hold teachers accountable, and to hold students accountable for learning. The most public misuse in recent years has been the high-stakes use of standardized test scores to promote and retain students and to determine whether students can graduate from high school.

Significance of the Study

Recently legislation in the United States, including the national No Child Left Behind Act and Georgia's A+ Reform Act, has begun to dictate that students, teachers, administrators, and schools as a whole be judged on the basis of very limited test results. High stakes have been attached to these tests, including funding of entire school systems. According to the No Child Left Behind Act, which became law in 2002, schools must make Adequate Yearly Progress in order to continue to receive funding. Specifically, the law attempts "to ensure that all children have a fair, equal, and significant opportunity to obtain a high quality education" (United States

Department of Education, 2002, p. 15). The law requires that it be demonstrated that students have achieved this equal opportunity by student performance on "high quality academic assessments" (United States Department of Education, 2002, p. 15). Further, the law holds local educational agencies (i.e., local boards of education) responsible for "turning around" those low-performing schools so that students all will have access to a high quality education. Students attending Title I schools that do not make Adequate Yearly Progress are to be provided with an alternative school within a school system that they may attend in order to get a better education. Transportation must be provided at the expense of the school system. If every school in a given system is a failing school, then up to 7 systems may pool their resources in order to provide alternatives for their students. In addition to providing alternative schools for students to attend, after school tutoring is another option that schools must provide in order to help students who perform poorly on state assessments.

One caveat of No Child Left Behind is that each state is left to create "high quality academic assessments,. . . curriculum, and instructional materials [that] are aligned with challenging State academic standards" (United States Department of Education, 2002, p. 15). Even though former Georgia Governor Roy Barnes and the Georgia State Assembly enacted the A+ Education Reform Act prior to the enforcement of No Child Left Behind, similar motivation for educational reform spurred both pieces of legislation. In 2000, educators in the state of Georgia were outraged when Governor Roy Barnes enacted legislation that was perceived essentially to undermine their work. This legislation also called for increased accountability for teachers, students, and schools based on standardized test scores. So incensed were teachers in the state that many attribute Barnes' failure to be elected to a second term to educators who actively lobbied against him because of the enactment of the law.

It is important to note, however, that these two pieces of legislation and others like them have not created the push for testing that exists in American society today. The outcry for accountability through testing has existed for decades and has been fed by the fact that the number of state departments of education increased three hundred percent from 1957 to 1986 and that the number of staffers for policymakers more than doubled from 1968 to 1979 (Firestone and Schorr, 2004). Standardized testing has been the tool that these departments have used to justify their existence because it is a concrete way to measure academic progress (or the lack thereof). This business-minded way to create more efficient ways of monitoring progress and hold schools accountable has come from people outside education. But it has only lately been that this outcry has become law.

Increasingly schools are held accountable for how students perform on a test. The flaws in this practice are many. Primarily, students, teachers, principals, and school districts are judged based on how students perform on a given test on a given day. The results are seen as the only method to determine whether or not a student is getting a quality education. That is to say, the lone test score determines if teachers and principals are doing their jobs. Yet, reliance on a single indicator has long been considered problematic. For example, in his *Basic Principles of Curriculum and Instruction* (1949), Tyler writes, "evaluation must appraise the behavior of students since it is change in these behaviors which is sought in education" (p. 106). He goes on to state that "evaluation must involve more than a single appraisal at any one time since to see whether change has taken place, it is necessary to make an appraisal at an early point and other appraisals at later points to identify changes that may be occurring" (p. 106). Further, Tyler writes that "there are a great many other kinds of desired behaviors which represent educational objectives that are not easily appraised by paper and pencil devices" (p. 107). Despite long-time

recognition of the limitations of relying on a single score, politicians in the United States continue to use single test scores as a means to categorize schools as low-performing.

The face validity of assessment instruments also becomes an issue. Face validity, according to James Popham (1988), is "a way of describing whether a test appeared (on the basis of visual inspection) to measure what it was supposed to" (p. 122). Further, Charles Mosier (1966) writes that "it is highly desirable" that a test has the "appearance of practicality" (p. 112). Unfortunately, as Cureton (1951) puts it, "A test is face-valid if it looks valid—particularly if it looks valid to laymen" (p. 672). These laymen, namely politicians and the general public, look at an assessment and see terminology that matches a particular curriculum. This leads them to the faulty assumption that the test necessarily measures that curriculum. In actuality, the measurement may measure the curriculum or it may not. Reliance on this kind of validity is seen by educational measurement experts as superficial and "is not a *Standards*-approved form of validity evidence and should, therefore, be employed by measurement people" (Popham, 1990, p. 97). Indeed, Remmers and Gage (1955) agree that it is "*not* permissible and may be very wide of the truth" (p. 124). Popham (1990) goes on to state, "Indeed, it was to extinguish the proliferation of such expressions as face validity that the *Standards* were originally produced (p. 97). The *Standards* to which Popham refers are the *Standards for Educational and Psychological Testing* as prepared by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education in 1985 (Popham, 1990). Cureton (1951) asserts, "So long as we realize that face validity is not logical relevance, no harm need result from attempts to make tests face-valid to increase their public acceptability, provided this does not result in weakening their logical or empirical relevances" (pp. 672-673). Often, standardized tests are seen by the public as a positive thing

because they appear to measure what they purport to measure; they are face-valid. The test is seen as a simple, quick, accurate method of evaluating students and schools. The flaw lies in the fact that face validity is not a scientifically or philosophically sound manner of determining the legitimacy of an assessment.

The public sees these test scores when they are printed in the local newspaper, and it becomes a contest to see which schools have outperformed the others. In fact, it is based on test scores that parents often make decisions regarding where they will purchase a home. Any self-respecting, successful real estate agent can tell a potential client the scores of the schools in the area. She can sell a house simply on the basis of the test scores of the neighborhood schools. After all, according to public opinion, test scores demonstrate just how well the school is teaching its pupils.

As a result of the public's desperate desire for such information, states are required to publish school report cards. In Georgia, the mandate came originally from House Bill 1187, otherwise known as the A+ Education Reform Act of 2000, prior to the requirement from *No Child Left Behind*. The report card notifies parents about how well or how poorly the schools that their children attend are doing. According to the Georgia Department of Education, the reports are "to ensure that a uniform standard is used throughout the state to determine needs improvement status of Title I schools" (Georgia Department of Education, 2002). According to the Georgia Department of Education, what the reports really reflect is some student and personnel demographic data and, for elementary schools, the test results for the Georgia Criterion-Referenced Competency Test for Grades 1 through 5 and the 5th Grade Writing Assessment. For middle schools, the reports give the disaggregated test results for the Georgia Criterion-Referenced Competency Test at grades 6 through 8 as well as the Middle Grades

Writing Assessment, which is an eighth grade assessment. The graphs presented on the reports demonstrate the students who did not meet the standard, who met the standard, and who exceeded the standard. At the high school level, the tests included are the Georgia High School Graduation Test, the Scholastic Aptitude Test, Advanced Placement Exams, and American College Testing. Also thrown in for good measure on high school report cards are graduate data, how many students earned the Hope Scholarship, and how many students entered public two-year or four-year colleges versus technical or adult programs. Additionally, the number of students requiring learning support (most widely known as remedial classes in college) is reported (2003). Schools are judged and ranked according to the number of students who do not meet the standard. There is no consideration given to any of the exigent factors that may result in lower scores for a particular population.

For example, schools with higher numbers of poor and minority students do not perform as well on most standardized tests as those schools with students who are white and from affluent families. In fact, African-American students score an average of 94 points less on the verbal section of the SAT and 104 points less on the math section than do white students (Lewis, 2000). The children of well-educated parents do better on standardized tests than those children whose parents did not graduate high school. These facts do not reflect the lack of student ability or the lack of good teachers in a school, but they can be a reflection of the bias in testing or the values of a given sub-culture or community. In rural Georgia where agribusiness is still the dominant force in the local economy, having a college degree is not as important as it is in suburban Atlanta. Furthermore, tests tend to favor students who live in suburban areas rather than rural ones and tend to favor white, affluent students over poor, minority students. Assessment expert Popham (2000) asserted that "If a child is raised in an affluent family, the odds are that the child

will be more likely to answer such test items correctly than if the child had been raised in a less affluent family" (p. 55). This type of socioeconomic bias makes the practice of using test scores to determine the fitness of school personnel baseless. Popham asserted, "When what's being measured is a student's socioeconomic status, not what the student has been taught in school, then it's clearly inappropriate to use the results of standardized achievement test as indicators of educational quality" (p. 58).

Most importantly, what looking at test scores leaves out is the individuality of students and school communities. Students and schools cannot be justly evaluated on the performance of a single day or week in the school year. However, each time that a score report is generated by a test publisher, curriculum decisions, such as whether or not a student will receive remedial classes or after-school tutoring and whether or not a student will be promoted to the next grade, graduate from high school, or go on to college, are made based on that student's performance on a single test on a single day. This snapshot of the student's educational progress is used to determine the ability of the student and the fitness of his school.

Purpose of the Study

The purpose of this study is to document and explain the appropriate and inappropriate uses of standardized test results as identified in the educational measurement literature in the United States from its origins in the early twentieth century to the present in order to inform contemporary educational policy and practice. This study expects to contribute to an awareness of the history of standardized testing protocol, to describe the misuse of standardized test scores in the United States, and to assist professional educators, policy makers, politicians, and the public in determining how to use this information to make better curriculum choices for students.

Scope of the Study

In order to attain the stated purpose of the study, the researcher will

1. Provide a context of the social and political climate that has generated the rise of the prominence of standardized testing in the United States and in the changes in the uses of those test scores.
2. Describe the history of the development of standardized test protocols of national tests such as California Achievement Tests, the Comprehensive Tests of Basic Skills, the Iowa Test of Basic Skills, Metropolitan Achievement Tests, Stanford Achievement Tests, the Scholastic Aptitude Test, and the National Assessment for Educational Progress.
3. Describe legislation and public policy, such as No Child Left Behind and the A+ Reform Act, that have dictated the use of standardized test scores deemed inappropriate by test publishers and scholars.
4. Report findings, conclusions, and recommendations for appropriate ways to evaluate student learning.

Research Questions

This study endeavors to answer the following questions:

1. What were the social and political factors that led to the rise of the standardized testing movement?
2. What are the appropriate uses of standardized test scores as identified by educational measurement literature in the United States?
3. What are the implications for contemporary educational policy?

Methodology

This study documents and explains the history of test protocol of selected major national standardized achievement tests and the theory that abounds concerning the uses of test scores. Kaestle (1988) writes that the field of educational history is similar to the field of historical research in that the generalizations are "the result of an interaction between fragmentary evidence and the values and experiences of the historian" (p. 61). This is essential to remember, Kaestle (1988) asserts, because current educational policies arise from the interpretations of educational history. Because of the different perspectives and interpretations, the same historical information can be used to argue two very different policies, as in the case of what is considered appropriate versus inappropriate use of standardized test results. Additionally, as Tanner and Tanner (1980) claim, if the field of curriculum theory ignores the rich history of curriculum, then perspective on contemporary problems is lost. Since standardized testing is so closely linked with curriculum decisions that are made for students, then it stands to reason, under the Tanners' argument, that an historical look at standardized testing will aid in gaining new insight that professional educators, politicians, policy makers, and the general public can utilize.

Fraenkel and Wallen (1996) assert that historical research can be used so that people "may learn from past failures and successes" (p. 495). Additionally, Fraenkel and Wallen write that if people learn about how things were done in the past, then a determination about the applicability to "present day problems and concerns" can be made (p. 495). Marius (1999) corroborates this when he writes that historical writing is an effort "to answer questions about origins, happenings, and consequences" (p. 2). As the last chapter of this study will assert, the issues surrounding the use of standardized test scores that are currently faced by professional educators can be addressed through knowledge of the past, including how scores have been used

throughout the twentieth century and how theorists and practitioners in the field of educational measurement believe that those scores should be used. More pointedly, this study will be an historical narrative that seeks to proffer "meaningful understandings" (Davis, 1991, p. 78). While the actual events that comprise the history of standardized testing will be chronicled, it will be the interpretation of what those events have meant to curriculum reform that will be of most significance to the researcher. Laurel Tanner (1983) writes, "To ignore our experience and retrace our false steps is to court almost certain waste and disaster" (p. 39).

As befits the nature of this study, the methods that are used in historical research will be used to explore this topic. Fraenkel and Wallen (1996) identify four steps to historical research (p. 497).

> These include defining the problem or question to be investigated . . .; locating relevant sources of historical information; summarizing and evaluating the information . . .; and presenting and interpreting this information as it relates to the problem or question that originated the study.

According to Marius (1999), historians want "to know what events mean, why they were important to what came afterwards, why we still talk about them" (p. 1). Because the field of standardized testing has such a rich history, it is necessary to determine what happened, why it happened, and what educators today can learn from those events. In order to achieve this goal, obviously the topic of research and the information that is relevant to reaching viable conclusions must be narrowed into a manageable focus. In deciding what information should remain and what should be discarded, Marius (1999) writes, "we try to sort [it] all out and arrive at the story that is most plausible to us" (p. 29). The five W's (who, what, where, when, and why) must be

answered in a fashion that does not exclude relevant information but also eliminates extraneous information that only clouds the issue. Marius writes that "journalistic questions help keep our eyes on this or that important thread so we can see how it contributes to the whole" (p. 33).

Marius (1999) also writes that "[h]istory includes data—evidence" (p. 4). For the purposes of this study, it is also important to note Marius' admonition to use "texts written by those you write about" to give "your own work authority" (p. 88). In other words, it is necessary to use primary sources when available to conduct historical research. It is the use of primary sources that gives credence to the work of the researcher. To give this study the authority to make the conclusions in Chapter V, past editions of the National Assessment for Educational Progress, the Scholastic Aptitude Test, the Iowa Test of Basic Skills, the California Achievement Tests, the Comprehensive Tests of Basic Skills, the Metropolitan Achievement Tests, and the Stanford Achievement Tests were studied. Specifically, the test protocols, which outline the appropriate uses for the test results, were explored. Some test companies refer to such documents as Interpretive Guides (Riverside Publishing and Iowa Testing Programs) while others call them Technical Handbooks (Educational Testing Service and College Entrance Examination Board).

In order to conduct this research, the researcher began by searching the Main Library at the University of Georgia (UGA) for such documents. Additionally, the Main Library at UGA was the source for many texts written by scholars in the field of educational measurement. To find these sources of information, the key words "educational measurement," "educational assessment," "standardized testing," and the names of each specific assessment that are the focus of this work were used as search terms in the library catalog. This included periodicals, texts, and Internet searches. Similarly, the search terms were used in several academic databases to

glean relevant information. The researcher utilized the services of a research librarian of the University of Georgia in order to search worldwide databases such as WorldCat. This effort resulted in several interlibrary loans of pertinent texts for the study. Finally, an Internet search of the topic was conducted. Many of the test publishing companies have detailed histories of their companies posted online, and this information was helpful in creating the timeline that later determined the publishing dates for the primary sources of information that were needed from the testing companies.

To get the primary sources of the test materials that identified the test authors' and publishers' purposes and intended uses for test results, the researcher tried a variety of techniques. The first step was to contact the National Center for Educational Statistics via email. The response from the organization was quick, and the resources that were necessary to the study were provided. The information from the National Center for Educational Statistics was available through the Internet and through various print sources at the Main Library at the University of Georgia.

The researcher also contacted Riverside Publishing, CTB McGraw-Hill, the College Board, and Harcourt Assessments via telephone and email, using customer service contact information. Emails, written letters, and phone calls to the companies went unanswered. Finally, letters written directly to the presidents of Riverside Publishing, CTB McGraw-Hill, and the College Board were successful. Within days, the acknowledgment of the correspondence began.

Riverside Publishing responded by sending Interpretive Guides dating from 1993 to the present. CTB McGraw-Hill's senior communications director responded by sending photocopies of pertinent information including technical handbooks for the California Achievement Test and

the Comprehensive Test of Basic Skills. In addition, a letter to the Dean of the College of Education at the University of Iowa resulted in communication from Dr. Stephen Dunbar, the Director of Iowa Testing Programs.

The assistant to the president of College Board provided a contact at Educational Testing Services. Once the Brigham Library, which houses the archives at Educational Testing Services, was contacted, an application was completed requesting permission to examine manuscripts. The completed application was approved in a matter of days, and a visit to the Brigham Library in Princeton, New Jersey, was scheduled. The researcher visited the Brigham Library, and an archivist was available for the day to assist her. Primary sources were in the form of microfiche, microfilm, and print. Photocopies of the materials were prepared for later analysis.

When even the letter to Harcourt Assessments president Michael Hanson went unanswered, the researcher contacted their Public Relations Department by phone. The request was forwarded to the appropriate person in research inquiry, and the researcher was contacted by phone from the office of a Permissions Analyst for the company. The researcher was then contacted by email with the appropriate citations that would be necessary for this document when referring to the Metropolitan Achievement Test and the Stanford Achievement Test. Materials on both assessments followed at a later date.

Several attempts to contact the Iowa Testing Programs for information were not acknowledged.

In order to interpret the literature in the field of educational measurement, the documents written by the scholars in this field will be analyzed. Kaestle (1988) stresses the integral part that theory plays in writing about history. In order to accurately portray the history of standardized testing, it will be necessary to look at the theories about testing that have prevailed during the

twentieth century. The primary sources that avow such theory include the work of Popham (1971, 1978, and 2001), Ebel (1977), and Kohn (2000). Additionally, the textbooks that have been written in the field of educational measurement will also prove useful in defining what authors deem appropriate and inappropriate uses of test scores. Textbooks from scholars such as Popham (1988 and 1990), Remmers and Gage (1955 and 1965), Lindquist (1951), Thorndike (1927), and Judd (1927) will be examined. As Fraenkel and Wallen (1996) point out, it is important to remember that, while generalizations are often a necessary part of historical research, one can increase the credibility of the generalizations by drawing from a large base of information. Therefore, the use of secondary sources will not be discounted altogether, but rather they will be used to provide an added perspective of the scholars in the field of educational measurement. Authors such as McNeil (2000), Popham (1974), Lagemann (2000), and Chase and Ludlow (1966) will be used. Additionally, relevant essays, books, and journal articles that were found in the University of Georgia Library, as well as through Georgia Library Learning Online and other Internet databases will be used.

The purposes for administering the tests, as explicitly stated by the test creators themselves, will be compared to current legislation, including the No Child Left Behind Act and the A+ Reform Act. This comparison will look closely at the purported use of test scores that is dictated by such legislation versus the intended use of the scores. Additionally, various speeches, press releases, and other comments by lawmakers are vital sources of information as to how public policy has developed.

Assumptions

In this study, the following assumptions are recognized:

1. The primary source documents examined for the purposes of this study are authentic.

2. The historical perspective on the appropriate and inappropriate uses of standardized test scores can inform current practice.

Limitations of the Study

The following limitations apply to this study:

1. This study included secondary sources when primary sources could not be accessed.
2. This historical study was limited by the accessibility of test protocols for the National Assessment for Educational Progress, the Scholastic Aptitude Test, the Iowa Test of Basic Skills, the California Achievement Tests, the Comprehensive Tests of Basic Skills, the Metropolitan Achievement Tests, and the Stanford Achievement Tests.

Some additional limitations to the study included the fact that test descriptions that included appropriate uses for each year that the SAT was administered were not available, although such documents were available for each year from 1956 through 1971. Additionally, the annual handbook was not available for every year, but the editions for 1948 and 1949 as well as technical handbooks from the 1970s were accessible. Score Use and Interpretation Manuals were available for 1926 (called the SAT School Manual in that year), 1959, 1960, and 1961.

The information on the intended uses of the NAEP were gleaned from various scholars and participants in the development of the NAEP as opposed to primary source documents similar to the Iowa Tests' Interpretive Guides.

Definitions of Terms

1. Criterion Referenced Tests (CRT): According to Gerald Bracey (1998), a criterion-referenced test is scored "in relation to a clearly specified set of behaviors" (p. 24). The levels of accomplishment are described, and students are evaluated by comparing the

students' scores against those descriptions. Susan Abbot (1997) describes this as measuring "development of a particular skill in relationship to absolute mastery" (p. 10).

2. Norm Referenced Tests: According to Gerald Bracey (1998), "a norm-referenced test is a standardized test with norms" (p. 19). A norm is a percentile rank of 50th that is given to the median score on a standardized test. Tests that are considered nationally normed are given to students all over the United States. Based on the results of that testing, students are assigned ranks depending on whether they score in the upper half (above the 50th percentile or above grade level) or the lower half (below the 50th percentile or below grade level).
3. High Stakes Tests: This term, coined by George Madaus, is used describe tests that are used to make decisions about students regarding grade placement, promotion, retention, tracking, and graduation (McNeil 2000). Madaus also describes high stakes tests as those that are tied to funding for schools or school districts, giving teachers merit pay, accreditation of a school or school district, or "placement of a school system into 'educational receivorship'" (1988, p. 30).
4. Accountability: Linda M. McNeil (2000) describes accountability as the use of students' individual and aggregate standardized test scores "as indirect measures of teachers' work, principals' 'performance', and even the overall quality of the school" (p. 6).
5. Psychological Testing: These tests are also known as intelligence tests. According to Paul Chapman (1988), the purpose of these tests is to measure the innate ability (or Intelligence Quotient—IQ) of a person and even to sort and track students.
6. Achievement Testing: According to Susan Abbott (1997), achievement tests "measure development or general achievement in one particular content area" (p. 10).

7. Standardized Testing: According to Gerald Bracey (1998), tests are considered standardized for four reasons. First, the format of every question is the same for every student. Secondly, every student receives the same instructions for taking the test. Next, the time permitted for each student to take the test is the same. Lastly, each question that the student is expected to answer is the same and has the same correct answer. Most often, a standardized test is given in multiple-choice format where there is only one correct answer. According to Robert L. Ebel (1977), an additional sameness is what ensures the test score means the same for each student tested.

Organization of the Report

Chapter 1 establishes an overview of the study, including background of the problem, significance of the study, statement of purpose, research questions, scope of the study, methodology, assumptions, limitations, definitions of terms, and organization of the report. Chapter 2 includes the origins and history of standardized testing in the United States and how the purposes of such testing have changed over time. This chapter also includes research on school reform and how it has shaped the face of educational testing in the United States. Chapter 3 explores the history of each of the seven assessments that are the focus of this research: National Assessment for Educational Progress, the Scholastic Aptitude Test, the Iowa Test of Basic Skills, the California Achievement Tests, the Comprehensive Tests of Basic Skills, the Metropolitan Achievement Tests, and the Stanford Achievement Tests. Chapter 4 documents the history of appropriate and inappropriate uses of results of nationally-used standardized tests as established by the educational measurement literature. Chapter 5 details the uses that are determined as appropriate or inappropriate by the test publishers and authors of the seven identified assessments. The final chapter summarizes the study, revealing the findings and

implications of the study to advance recommendations for educators and public policymakers. Suggestions for further research are also included in the final chapter.

Chapter 2

THE ORIGINS AND HISTORY OF STANDARDIZED TESTING IN THE UNITED STATES

There were three early events that led to modern psychological and educational testing. Dubois (1970) identified those three as follows: Civil Service Exams, the studies of individual differences in Europe and America, and the assessment of academic achievement in universities and schools (accountability). This chapter will explore how the history of psychological and educational testing has evolved during the twentieth century. This exploration of the development of standardized testing in the United States will lay the foundation for understanding of current educational practice and public policy. The philosophies and practices of both the theorists and practitioners who dominated the field of educational measurement throughout the century will be highlighted, and the evolving theories as well as the steadfast perspectives will be investigated. This examination will include a discussion of the need for testing to sort the flood of diversified students in schools due to compulsory attendance laws and will extend to encompass the changing social environment that allowed testing to become a stronghold in American education. Additionally, the advent of testing as big business in the United States will be discussed along with the use of such standardized tests as a way to hold schools accountable.

Civil Service Exams

According to Wardrop (1976), the concept of testing itself goes back to 2200 B. C. when the emperor of China began using civil service examinations to determine the fitness of public officials. Every three years, the government administered the examinations, and, after three

examinations, officials were either promoted or released from service based on the results. The exams initially consisted of tests of the arts such as horsemanship, writing, and arithmetic; however, by 1370 A. D., the holders of high public offices were required to know and to interpret Confucius (Dubois, 1970). It is important to note that the Chinese used these exams because they had no public schools or universities to ensure that the general populace would have the requisite knowledge and skills to hold public office successfully (Dubois, 1970).

Dubois (1970) also emphasized that, in 1791, France's Voltaire and Quesnay advocated the use of the Chinese examinations, and their appearance in France was generally considered a reform effort. However, Napoleon did away with the exams while the rest of Europe began to rely on them more heavily. In fact, the widespread use of exams in other parts of Europe caused the government of France to re-instate such civil service exams. In addition to the use of these exams in France, the English government began to use them as a way to find qualified participants in the Indian civil service. Much later, in 1883, civil service exams were established in the United States.

In 1883, the United States Civil Service Commission came into being (Wardrop, 1976). There were approximately 14,000 jobs that "were subject to competition. On June 30, 1922 about 420,000 persons held positions in the civil service subject to competitive examination" (Deming, 1923, p. 198). Over 250,000 people were taking exams yearly during the early part of the twentieth century (Deming, 1923). The Commission defended their "careful and complete system of rating [that] insures a fair and impartial judgment of the relative merits of the applicants" (Deming, 1923, p. 198). It is key to take notice that, according to Dubois (1970), even in 1883, the exams were considered only as a part of the whole picture. Applicants were asked to submit work samples of previous work that they had done in the field of desired

employment (Gardener, 1923). And it appeared that the Commission was giving early versions of performance assessments. "Wherever practicable, actual tests are given in the work to be performed. For example, applicants for stenographic positions must be examined in practical stenographic work" (Gardener, 1923, p. 199). The government made a final determination of the civil servant's suitability for the position by looking at a combination of criteria: the results of the examination in conjunction with the application he submitted and his job performance during the first six months. This is contrary to the current trend in education reform which calls for test scores to the sole determining factor of a student's knowledge and of the effectiveness of a school.

Studies of Individual Differences in Europe and America

During the nineteenth century, the most influential portion of the history of educational testing began with Alfred Binet in France, with Sir Frances Galton in England, and with James McKeen Cattell in the United States (Wardrop, 1976). It was Galton who introduced a psychology based on the differences found in individuals. In fact, he was the first person to develop tests of psychological function (Dubois, 1970).

Sir Frances Galton was a man of many talents and interests. In addition to his contributions to the field of psychological testing, he accomplished "firsts" in other fields such as meteorology, biology, and anthropology. For example, he was the first person to publish weather maps and started the study that eventually led to the use of fingerprints as a method of identifying individuals (Dubois, 1970). He was also the cousin of famed evolutionist Charles Darwin and a self-proclaimed eugenicist (Lemann, 2000). Galton believed in "the systematic biological improvement of the human race" (Sokal, 1987, p. 26). The role that eugenics plays in the testing movement will be discussed in more detail later in this chapter.

In the pursuit of his interest in individual differences, Galton administered many different kinds of tests to his subjects that were physical in nature, such as height, weight, and hand steadiness, but which led to more behavioral testing (Wardrop, 1976). Other testing that Galton did was comprised of test of discrimination of weights, visual images, and audible sounds (Dubois, 1970). He used the data from the testing of his subjects to create "tables of percentile norms by sex, for several physical and behavioral characteristics" (Dubois, 1970, p. 14).

In 1892, Alfred Binet began to introduce and administer tests of intellectual functioning at the Sorbonne (Wardrop, 1976). In 1905, along with T. Simon, he developed the first modern psychological test, "the prototype of all subsequent mental scales" (Dubois, 1970, p. 36). On commission from the French government, the scale was originally developed to assist in the education of students who were intellectually disabled (Wardrop, 1976). In fact, his work led him to advocate that the education of such students take place in "special schools or in special classes in regular schools" (Dubois, 1970, p. 34). Binet used mental tests "to detect differences among children in such things as recall, moral judgment, and mental addition" (Chapman, 1988, p. 19). In France in 1904, tests were used to segregate students who were "not making adequate progress" into separate classes (Chapman, 1988, p. 19).

According to Dubois (1970), memory span was one of the first concepts with which Binet worked that later became a part of his intelligence tests. Binet included two types of tasks on his tests. One was the constant task. In the constant task, the result actually became the measure. For example, he showed a child a list of items for five seconds and then asked him to remember as many items from the list as possible. The other kind of task was the variable task in which the measurement was adjusted to the ability of the student. An example of this kind of

task was to measure the longest series of objects (say, numbers) that a subject could repeat without error.

Binet found that the scores he obtained from such testing of both intellectually talented and intellectually disabled students yielded a "bell curve" distribution of scores and seemed to correlate to grades in school (Wardrop, 1976). At about this time, two psychiatrists, Blin and Damaye, were testing the mentally disabled as well, and Binet borrowed their idea of a total score on a test "to reflect gradations in intelligence" (Dubois, 1970, p. 33). However, Binet was also highly critical of their test questions that did not require the subjects to think.

Dubois (1970) maintained that, in 1905, Binet published the first scale that compiled more than thirty tests with which he had been working for a number of years. "Using the scale, Binet could determine a child's 'mental age,' his relative intellectual development, and through 'age norms' he could compare the child's ability with that of others the same age (Chapman, 1988, pp. 19-20). While he made many improvements to the items over the years until his death in 1911, many of the questions that were on the 1905 scale are still on the third edition of the Stanford-Binet scale published in 1960. But in 1908, a new publication of the improved test included fifty-eight separate subtests. With this publication came a determination of the subject's mental age. Mental age was measured to be the level at which students passed all of the tests but one at a particular level, adding one year of mental age for passing five tests above that level and adding two years of mental age for passing ten tests above that level. Through Binet, yet another admonition was issued not to use the tests alone to determine a student's mental age, ability, and achievement. Binet "pointed out that an a priori system of measurement probably would not fit the great variety of expressions of intelligence. . ." (Dubois, 1970, p. 32). Additionally, Binet and Simon stated, "'The results of our examination have no value if

separated from all comment; they must be interpreted'" (Dubois, 1970, p. 39). This is understood to mean that Binet knew that to take a score in isolation was to leave out considerations that would assist in determining the true mental ability or achievement of an examinee.

Dubois (1970) reported that, by the time Binet died in 1911, he had published work on the following issues: correlations between school performance and mental ability, the effect that retesting has on test scores, teacher methods in judging intelligence, and correlations between the socioeconomic status of the subject and measured intelligence. However, Binet did not only leave behind the actual tests and these analyses of the results, but also he left the world with important guidelines for testing that educators still adhere to today. He encouraged the practice of having just one examiner and one examinee and of the importance of the examiner's rapport with the student. Binet also admonished test administrators to start testing at a level that is high enough not to insult the student and to finish testing before the student gets too tired to do her best or loses interest. Finally, Binet warned about keeping meticulously accurate test records. Binet maintained that all of these procedures were (and still are today) crucial to obtaining accurate results from evaluations. Binet simply thought of himself as providing a way "to identify slow learners so that they could be given special help in school" (Lemann, 2000, p. 17). Binet and Simon (1916) stated, "But we are of the opinion that the most valuable use of our scale will not be its application to the normal pupils, but rather to those of inferior grades of intelligence" (p. 263). The desire to help those students with a course of education that would be most suited to their needs was of the utmost importance to Binet's work.

"It was the work of Alfred Binet in France that had the most direct influence on the development of intelligence tests in America" (Chapman, 1988, p. 19). Many American

psychological testing pioneers adapted the work of Binet. "Psychologists saw in intelligence tests a way to improve schools and to enhance the reputation of their science" (Chapman, 1988, p. 18). In 1908 and 1910, Goddard published modified versions of Binet's work as did Whipple in 1910 (Dubois, 1970). Yerkes, Bridge, and Hardwick published variations of Binet's tests in 1915, and Yerkes followed up with another publication in 1923. Even though "testing was never his primary goal or interest" (Reed, 1987, p. 79), Yerkes, too, was greatly influenced by Darwin and the idea of eugenics. His work during World War I with examinations in the Army (the Army Alpha and Army Beta) earned Yerkes a place in the burgeoning world of testing. At the end of World War I, "Yerkes's work in developing diagnostic mental tests attracted national attention" (Reed, 1987, p. 81).

The most accepted American modification of Binet's work was by Lewis M. Terman, a professor at Stanford University. It was Terman and Edward Thorndike who began to see Binet's work as so much more than Binet ever did. Terman and Thorndike are the ones who were "advocates of the widest possible use of IQ testing by American educators, so that students could be assessed, sorted, and taught in accordance with their capabilities" (Lemann, 2000, p.18). Terman "saw immediate application for the tests in the diagnosis of individual problems, for they offered a 'more reliable and more enlightening estimate of the child's intelligence than most teachers can offer after a year of daily contact in the schoolroom'" (Chapman, 1988, p. 25). Additionally, Terman believed that widespread use of intelligence testing would be beneficial. "Not only in the case of retarded or exceptionally bright children, but with many others also, intelligence tests can aid in correctly placing the child in school" (Terman, 1916, p. 16).

But Terman thought that the greatest use of the scores would be for "atypical children in the public schools" (Chapman, 1988, p. 25). Terman mainly thought that the tests could be used

for vocational guidance, particularly for those students who were mentally "retarded" (Chapman, 1988, p. 26). "Regarding the nature of intelligence, he argued that his test constituted a valid measure of intelligence, that the IQ was constant, and that it was greatly influenced by heredity" (Chapman, 1988, p. 28). Terman also thought that "identifying 'feebleminded' individuals" would help alleviate social ills such as poverty, crime, and delinquency" (Chapman, 1988, p. 32). Terman's belief that the IQ did not change and that it was largely the result of heredity fit in somewhat with the eugenics movement that had taken hold of the field of educational measurement at this time. But Terman also thought that genetics played only a portion of the role in determining intellectual ability. He believed that ". . .without such tests we cannot know to what extent a child's mental performances are determined by environment and to what extent by heredity" (Terman, 1916, p. 19). He believed that the factors which influenced intelligence could be determined by the use of intelligence tests (Terman, 1916). Thorndike commented "that intelligence can be increased only if eugenics encourages the bright to have more children and the dull to have fewer" (Joncich, 1968, p. 322).

In the 1920s, "Terman was making great strides toward accomplishing his goal of reorganizing the schools through testing" (Minton, 1988, p. 100). He was working on this through touting his belief that teachers were the main factors in the use of testing in schools. Terman (1919) was adamant that teachers "must learn to use tests" (p. 291). He believed fully that tests would be valuable in assisting with vocational guidance in terms of whether or not a particular vocation was "compatible with the general mental ability which an individual possesses" (Terman, 1919, p. 270). Terman insisted that teachers rate their students "for quality of school work, general intelligence, and two or three personal traits like dependability, social adaptability, conscientiousness, etc." (p. 301) in order to get a whole picture of the child in

conjunction with the results of an intelligence test. This was a very important distinction to be made. Even for Terman, tests were just one part of what should be considered when making decisions for a student.

Although he was criticized sharply from some corners, Terman was undeterred from building an alliance with the publishing industry. He saw this as a very important way to establish testing in the schools. "This alliance also proved helpful in fending off the testing critics" (Minton, 1988, p. 107). Since many of the publishing companies were also the publishers of the textbooks in use at the time, the public may have had the same confidence in their testing products that they had in their textbooks (D. P. Resnick, 1981). Terman was depending on that fact. Additionally, large scale publishing made the tests economically efficient for schools and school districts. Both the textbooks and the tests "met a need for national standards" (D. P. Resnick, 1981, p. 627).

While Alfred Binet was doing his groundbreaking work in France, an American, James McKeen Cattell was at Columbia University creating psychological tests for sight, hearing, taste, smell, and "mental time" (Wardrop, 1976, p. 8). "In an article written in 1890 Cattell coined the term 'mental test' and set the tone for America's practical application of the new measures. . . 'The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances'" (Chapman, 1988, p. 20). His "sensory and psychomotor tests," however, bore out no correlation to grades (Wardrop, 1976, p. 8). Cattell maintained that his battery of tests of physical reactions was psychological in nature and that "it is impossible to separate bodily from mental energy" (Dubois, 1970, p. 17). Cattell administered his tests to fifty freshmen at Columbia and to some women at Barnard, and, since there was no correlation between the students' grades and the

results of the testing, Cattell's work appeared to be a failure (Dubois, 1970). In fact, according to M. M. Sokal (1987), Cattell abandoned his own work in the early twentieth century because his tests gave no valuable information. And it was in 1917 that he was dismissed from Columbia University because of his unpopular political views. He "publicly opposed the use of the draft in World War I to obtain soldiers for combat service" (Harcourt Brace and Company, 1994, p. 59). Despite his views and the seeming failure of this work, it was what rose out of his work that made him successful.

Cattell had many faithful students who followed in his footsteps, one of whom was E. L. Thorndike (Wardrop, 1976). While Thorndike was at Columbia, he created tests of arithmetic, handwriting, spelling, language, and reading. In fact, it was during this time that Columbia University became the central point of development for the field of educational testing (Dubois, 1970). "As Thorndike remarked some years later, his early interest 'concentrated on research methods of measuring mental abilities,' but by 1909 he and his students focused increasingly on 'scales for use in measuring school achievement in reading, writing, drawing, composition, knowledge of history and the like'" (Chapman, 1988, p. 21).

While Galton, Binet, and Cattell were seeming main players during the early portion of the history of testing, they turn out to be very minor contributors. However, there is one attribute that they and a host of other European and American scholars shared: the main focus of their efforts was to measure intellect, and intellect is defined by Dubois (1970) as being the combination of sensation, attention, perception, association, and memory.

The New Phenomenon

In following the track that has led to the multimillion dollar industry of publishing, scoring, and reporting standardized tests, it is important to note that, prior to World War I, there

were two types of tests: individual intelligence tests and achievement tests (Wardrop, 1976). It was these examinations that led to the inception of the Army Alpha, which was the first large-scale use of intelligence tests (Dubois, 1970). The Army Beta, used for those who were illiterate or who did not speak English, also became popular for use in assessing the abilities of soldiers that were already in military service during World War I and was just the beginning for group testing. Yerkes saw intelligence tests as a way for the Army to classify recruits, and he was joined in his opinion by Terman (Chapman, 1988). In fact, Yoakum and Yerkes (1920) claimed that the use of Army tests "as partial basis for placement of soldiers; to the second as supplementary information for guidance in connection with training, or special treatment of men who give trouble; and to the third, as partial basis for recommendation for discharge, special examination, or medical treatment" (p. 47) was an appropriate use of the group testing. Additionally, Yoakum and Yerkes (192) declared that the exam "helps to reveal non-commissioned officer material and suitable candidates for officers' training camps. It also supplies partial basis for assignment of men to specified trades or occupations in the Army" (p. 47). During this time period, group testing was used for vocational and aviation testing as well.

During this time, an interesting phenomenon of growth was taking place in the United States. Between 1890 and 1917, the population of the United States doubled, primarily through immigration (Chapman, 1988). Judd (1933) claimed, "The increase in attendance on high schools is a result of a widespread popular demand for extension of the opportunity for free education above the elementary level" (p. 160). Chapman went on, "School enrollment increased by over 50 percent. Student went to schools with longer terms, and attended more school days" (1988, p. 41). Yet another factor, according to Judd (1933) and Chapman (1988), contributing to the boost of school attendance was the increase in compulsory attendance laws.

"School attendance was up markedly around this time, rising from 136 persons in 1,000 in 1904 to 152 persons in 1916" (Chapman, 1988, p. 43). By 1930, enrollment had soared from the 1900 figures of 700,000 to 4.8 million (Angus and Mirel, 1999). Not only did the shear numbers of student increase, but also the percentage of adolescents who enrolled in school increased. According to Angus and Mirel (1999), in 1890, only 5.6 percent of the population aged 14-17 was enrolled in school. By 1900, that percentage nearly doubled to 10.2 percent, and by 1930, that figure had reached a staggering 50.7 percent of the population aged 14-17 was enrolled in school (Angus and Mirel, 1999). In 1890, high school graduates represented only 3.5% of seventeen-year-olds, whereas in 1930 this figure was 29% (Angus and Mirel, 1999). Judd (1928) asserted, "The fact is that the schools are increasing so rapidly in the number of pupils enrolled that the generation which has completed its education is not able to supply teachers in adequate numbers" (p. 42). Valentine(1987) reported, "The continued explosive growth of public high schools—the number of graduates more than doubled from 240,000 in 1915 to 528,000 in 1925—created an ever-larger pool of college applicants" (p. 31).

Chapman (1988) explained ". . . [I]ntelligence tests were adopted in the public schools because the tests reflected widely shared values of the Progressive Era. University professors and school people alike saw the tests as a logical outgrowth of the progressive quest for efficiency, conservation, and order. The tests were welcomed by people who placed their trust in the authority of science and the expert" (p. 5).

This great rise in the number of students that needed to be educated as well as the great diversity in the students who were coming to school resulted in the uncertainty as to how to manage them all. One solution in order to adequately educate them was to sort students. The students at this time were not only increasing in number, but they were also diversified in ability

and in the aims of their education. The social background of students as well as their ability levels was varied. According to Angus and Mirel (1999), "for every 3 ninth-grade students from white collar families there were two students from blue collar families" (p. 33). Many sought an education not because they desired one but because they were required by law to attend. According to Angus and Mirel (1999), educators pushed for a differentiated curriculum to appeal to vocational training for the massive numbers of students coming to high school, but parents and students didn't take advantage of those opportunities. The overwhelming majority of classes were in traditional academic fields. Despite the differences in the backgrounds and abilities of students who were entering high school, "only 20% of total course enrollments were in the vocational subjects, and more than half of these were in the commercial field" (Angus and Mirel, 1999, p. 48). The differentiation in terms of curriculum was not very great, so tests became an integral part of schooling in order to sort students in their academic courses. This use of intelligence tests was, again, a simple and efficient way to manage the high numbers of students who did not seem to be as academically-oriented as those of past years. The reason for this is quite simple: compulsory attendance laws required all children to attend school regardless of their ability or their value of education. This meant that students might not be as capable in general as the overall population in schools up until this time, which was made up of students from higher socioeconomic conditions or those who were motivated to be in school by their own (or their family's) desire to get an education.

As Madaus and Kellaghan (1992) declared, "Faced with large numbers of diverse students being forced to attend school and with large numbers not doing well and being retained in grade (Ayres, 1909), acceptable achievement levels on tests were relativized" (p. 122). Though many educators of that time sought absolute test scores because they considered them

essential to the management of such large numbers of students, these results eluded test users. That is to say that the users of tests wanted to be able to determine that a student who scored at a certain level would receive a certain course of study. Instead, the tests yielded scores that were seen in the light of other factors such as motivation and desired vocation, and it was these factors that weighed heaviest in making curricular decisions for students.

In 1916, the first edition of the Stanford-Binet Intelligence Scale was published with revised editions issued in 1937 and in 1960 (Dubois, 1970). The 1916 version was the first to yield a test result to be called an "IQ," which was derived from William Stern's "mental quotient" (Dubois, 1970, p. 51). The first version was also the closest to Binet's original work in that only minor adjustments were made to the levels at which tests were placed. Only three of Binet's original tests were eliminated, and several of Binet's earliest tests were added to the test, along with some of Terman's own work (Dubois, 1970). These few changes were a testament to the quality and validity of the original work done by Binet and his reputation in the world of psychological testing. Virgil Dickson was one of Terman's students at Stanford. "Based on his first year's work, Dickson observed that 'standard tests, both psychological and pedagogical—group and individual—should be of great assistance in the classification of pupils according to ability and capacity to do work'" (Chapman, 1988, p. 56). Among other uses, Dickson used these tests to place first grade students. This fit the purpose for which standardized tests had become popular—utilitarian to deal with the booming population and the steady rise in the numbers of pupils.

It was around this time that Terman realized that his prescription to test "retarded" children was not working because "'[t]he number of school laggards has decreased but little, and their needs are almost as little provided for as before the campaign on their behalf began'"

(Chapman, 1988, p. 85). Terman thought that "'a reasonable homogeneity in the mental ability of pupils who are instructed together'" (p. 86) would be an efficient way to conduct schooling, and intelligence tests provide the information to classify children in this way. "Terman mapped out a plan that could be adopted by any school system. He proposed giving all children an individual intelligence test in the first grade" (Chapman, 1988, p. 87). Terman believed that a student's IQ would help determine "'what the child's future development will be'" (Chapman, 1988, p. 92).

While many of this time advocated the use of test scores to sort students and to make determinations about their futures, there were many others who questioned this practice. One such notable figure of the 1920s was William Chandler Bagley. ". . . [H]is critique of educational determinism and what he viewed as the inappropriate application of intelligence testing to the practical work of schooling largely was ignored by leading developers of intelligence tests and advocates of their use" (Null, 2003, p. 6). Bagley "attacked the intelligence testing movement because he thought it was based on bad science, because he viewed it as anti-democratic, and because he thought these new and incompletely formed theories were being applied hastily to educational practice" (Null, 2003, pp. 181-182). He even referred to the tests as "so-called intelligence tests" (Bagley, 1925, p. 5). And Bagley was also concerned that intelligence testing was antithetical to "his ideals of science, democracy, social service, and more importantly, education" (Null, 2003, p. 181). According to Null, Bagley asserted that the others in the testing movement wanted only to give few tests to all students in all schools without much thought to the individual. While he "recognized some value in their use as far as assisting student to make choices for future employment, . . . Bagley thought this insistence upon deduction ignored the particulars of individual students and individual schools. To Bagley, the

greatest threat in the practical application of these intelligence tests was to democracy and its precious institutions" (Null, 2003, pp. 181-182).

In 1922, Bagley began the "intraprofessional debate" (Chapman, 1988, p. 130) over testing . "Disclaiming any 'personal animus,' he warned that the 'present tendency' to increase the use of tests beyond a 'very restricted field' is 'fraught with educational and social dangers of so serious and far-reaching a character as to cause the gravest concern'" (Chapman, 1988, p. 130). Bagley asserted that using tests extensively would "undermine democracy by promoting an increasingly stratified society" (Chapman, 1988, p. 131).

At the least, according to Wardrop (1976), the IQ scores that are derived from standardized tests "reflected social-class bias" (p. 13). In fact, there were many who believed that standardized tests actually made it acceptable to discriminate against individuals based on their intelligence. Wardrop (1976) additionally made sinister notation of the fact that the majority of the members of society did not notice this for decades. His observation was followed with accounts of laws in twenty-one states that required sterilization because "the feebleminded were incapable of moral judgments" (Wardrop, 1976, p. 13). The idea of eugenics and standardized testing had become intertwined.

According to Lemann (2000), "True believers in IQ tests thought they should be given to all American school children, so that the high scorers could be plucked out and given the best schooling and the average and low scorers consigned to a briefer, more limited education" (p. 23). It was on this premise that the eugenicists took hold in American education. One of the leading eugenicists was Darwin's cousin, Galton. In his work, *Hereditary Genius*, Galton expressed his belief that intelligence was the most important characteristic and that whites were more intelligent than dark-skinned people. This ridiculous notion of his was accompanied by the

idea that the less intelligent people of the world were having more children than the more intelligent people and that this was bringing down the average IQ of the entire human race (Lemann, 2000). In fact, Galton invented the term eugenics as a way to describe "selective breeding techniques to improve the intelligence . . . of the human race (Lemann, 2000, p. 23).

Joining Galton were the likes of Carl Campbell Brigham (the author of the SAT), who also subscribed to the belief in eugenics (Lemann, 2000). These men and many others who felt that the white race, and intelligent white people in particular, were superior ranked Europeans from highest to lowest groups in this order: Nordics, Alpines, and Mediterraneans. In the United States, eugenicists felt that the influx of Mediterraneans as immigrants in to the country was diluting the pool of intelligence. While ". . . Americans did not publicly announce that they were sorting their children on the basis of IQ tests" (Lemann, 2000, p. 119), that was exactly what was happening all over the country.

Judd (1916) asserted, "If we consider individual cases, we find that there are some children who fail because of lack of native ability to do the work. . . They ought to be given other kinds of training which will reach their level" (p. 18). Judd (1916) contended that "[n]o school system can free itself entirely from the difficulties which are so clearly revealed by these tests and comparisons. The children in different schools differ one from another" (p. 58). The individual student, the individual school, and even the individual teacher had an impact on the performance of students on these tests.

In 1926, "a committee of outstanding educators, including Bagley, Bobbitt, Charters, Counts, Judd, Kelly, Kilpatrick, and Rugg, . . . argued that 'to serve a useful purpose, tests must be fitted to the requirements of the curriculum. . . .'" (Madaus and Kellaghan, 1992, p. 125). The committee "condemn[ed] the use of standardized tests by administrators to evaluate the product

of education" (Madaus and Kellaghan, 1992, p. 125). They thought that this would be a great misuse of tests and their results. The publication of The Twenty-Sixth Yearbook of the National Society for the Study of Education brought an official stance of many of the leading educational scholars of the period to testing. In its publication subtitled "The Foundations and Technique of Curriculum Instruction", there were three statements that addressed measuring the outcomes of instruction:

Measuring the Outcomes of Instruction

49. One of the most potent form of curriculum-control is measurement by means of uniform examinations and standardized tests. Teacher and pupils will inevitably work for the elements represented in the instruments by which their success is measured: therefore, it is of the utmost importance that changes in goals and methods be accompanied by the development and use of new tests and examinations corresponding in type to the advances made in the curriculum. To serve a useful purpose, tests must be fitted to the requirements of the curriculum and to the requirements of method. They must be determined by the purposes set up in the curriculum for the group of children being tested.
50. This Committee condemns emphatically the evaluation of the product of educational effort solely by means of subject-matter types of examinations now prevalent in state and local school systems. We have reference specifically to the rigid control over the school curriculum exercised by those administrative examinations which over-emphasize the memory of facts and principles and tend to neglect the more dynamic outcomes of instruction.

51. The foregoing statement is not to be construed as interfering in any way with tests of any character given intelligently for general scientific research. (p. 25).

These statements clearly advocated the use of examinations that are more closely aligned with a curriculum that is not solely about activities that promote the memorization of facts. This criticism of standardized testing came at a pivotal time when many respected educational theorists and practitioners found themselves at odds with each other. The importance of this particular section of the report was highlighted by Horn (1926): "This section [on testing] is, in the writer's opinion, one of the most important in the entire report. Progressive changes in curricula will be made by schools under considerable penalty if the success of such changes in the school system is measured by tests which do not take in to consideration the purposes for which the changes have been made" (p. 111).

Even those members of the committee that disagreed with parts of the statement all agreed with the sections on testing. The overarching theme was that testing and the progressive movement were not necessarily compatible unless the kinds of assessments that students were given to evaluate their learning (and ultimately the curriculum) were changed from the multiple-choice, standardized format of questions of fact that required little deep thinking or consideration of the individual. The kinds of assessments that the Progressives called for would have included those tasks that called for critical thinking, problem-solving, and analysis and application that was pertinent to everyday living.

And it was the misuse of the tests that concerned others like Walter Lippmann, who supported classifying students. Chapman (1988) declared the following:

In his first article, Lippmann had warned that because of unreasonable claims

> and faulty assumptions the tests were "in danger of gross perversion by muddleheaded and prejudiced men." In his final essay, "A Future for the Tests," he explored the possible key role in school for tests should the view become common that they measured hereditary intelligence. The tests and their makers would "occupy a position of power which no intellectual has held since the collapse of theocracy." The testers would become gatekeepers at the door to opportunity. (pp. 136-137).

But Terman "asserted that abuse was not 'one of the recognized rules of the game'" (Chapman, 1988, p. 137). Opponents like Trabue "pointed out the dangers that would stem from confusion about the nature of the tests, exaggerated confidence in their accuracy, deterministic applications of test results, and inappropriate use in the schools" (Chapman, 1988, p. 140).

Unfortunately, abuse of tests may have not been a recognized rule of the game, but it did become an unspoken rule. The use of test scores expanded greatly. According to Chapman (1988), the number one use of group intelligence test in elementary schools in 1925 was the "classification of pupils into homogeneous groups" (p. 156)—used by 64 % of cities in elementary schools, by 56% in junior high schools, and by 41% in high schools. The other uses were as follows (Chapman, 1988):

2. Supplementing teachers' estimates of pupils' ability
3. Diagnosis of cause of failure
4. Establishment of classes for subnormal children
5. Extra promotions
6. Comparison with other school systems
7. Admission to first grade of elementary school (pp. 156-157).

Many reports suggested that "test use was most extensive in the lower grades" (Chapman,

1988, p. 169).

According to Lemann (2000), in 1928, Brigham, who would later write the highly successful SAT, shocked his circle of friends by publicly denouncing his earlier statements that supported the eugenics movement. In fact, Terman invited Brigham to speak to a conference that was focused on the theory of superior intelligence, and Brigham declined. Brigham began to write that he did not want to do away with testing altogether; however, he thought that we should not put so much importance on the score and what it actually meant about intelligence. In a manner of speaking, Brigham was further validating the views expressed by other researchers before him, most notably Binet, that educators and the "powers that be" were putting entirely too much stock in what one test score can tell about an individual's mental ability.

There were proponents of the testing movement who wanted to be heard as well. Judd (1934) declared the following:

> Tests not only furnish a basis for the comparison of classes and individual pupils but are instruments of scientific study of important problems with which education must deal. They have what has been called 'diagnostic' value. Professor Ralph W. Tyler, of Ohio State University, gave a striking demonstration of the utility of tests in throwing light on problems of college education. . . It was found that knowledge of terminology and knowledge of facts were distributed among the students in much the same way. . . . Power of inference, on the other hand, showed a very different distribution from that shown by the other types of achievement measured. (pp. 233-235).

Judd claimed that mental testing let educators know "what may reasonably be expected of a given child" (p. 236). Judd believed that testing had its place in education to

highlight where the weaknesses of a particular student lie. Ralph Tyler, to whom Judd referred, continued to be a force in the field of educational measurement and will be discussed later in this chapter in regards to his work with the National Assessment for Educational Progress.

Educational testing received its big boost with the widespread use of the multiple choice test item, which became immensely popular when it was first used in the Army Alpha. The field gained even more ground with the rise of the "printed test of intelligence" (Dubois, 1970, p. 73). By the 1920s, the number of standardized tests that were approved and in use skyrocketed (Dubois, 1970). From 1925 to 1950, there were three main areas of growth for the industry, according to Wardrop (1976). They included creating new statistical procedures for analyzing and improving tests, creating faster way of scoring the tests and reporting the results, and creating the "institutionalization of testing in American society" (Wardrop, 1976, p.14). It was also during this time that the some of the most vocal detractors of schooling in America came forward. In 1925, Courtis first argued that educational testing reveals the inefficiency of schools (Dubois, 1970). Courtis was quoted by Dubois (1970) as stating the following:

> How great the inefficiency of public education really is few realize in spite of the repeated revelations of survey data. The results of Rice, the pioneer in the modern movement for exact comparative measurement, were received with open disbelief and ridicule; but they have been substantiated by survey after survey the country over. Today there can be no excuse for ignoring the fact that very few children profit as they should. . . . (p. 72).

Whether or not Courtis was looking at several measurements to determine this inefficiency or not, as Binet warned about doing, was uncertain. What was certain was that

Courtis was just the first in a very long line educators, businessmen, and politicians who would use standardized test scores as a means of indicting American schools.

In stark contrast to the eugenics movement, the 1930s brought an era of more democratic thought. James Bryant Conant, the president of Harvard at the time, wanted to follow an idea from Thomas Jefferson and create a "natural aristocracy" that was selected democratically from all walks of life but that was still the most highly intelligent group (Lemann, 2000, pp. 78-79). Conant retracted "his stance in utilizing standardized tests for gauging school achievement and determining college admissions. 'I recognize the multiplicity of examinations and am ready to support a proposal for more emphasis on school records and less emphasis on examinations,' stated Conant" (Tanner and Tanner, 1990, p. 339). Conant felt that intelligent people could be found in all socioeconomic groups and that such individuals should be educated so that they could fulfill their potential. Conant stated, "'Furthermore, education can inculcate the social and political ideals necessary for the development of a free and harmonious people operating an economic systems based on private ownership and the profit motive but committed to the ideals of social justice'" (Passow, 1977, p. 4). Conant also advocated the expansion of "vocational education—what is later expressed by Conant in terms of the 'acquisition of marketable skills'" (Passow, 1977, p. 6). He felt that this commitment to public education funded by the public would improve American democracy.

The Business of Standardized Testing

The pioneering stage of the development of standardized testing ended in the United States during the 1930s, and test validity became the focus of the field (Dubois, 1970). In this time period, researchers and test writers hailed only two ways to ensure test validity: to test and retest and to correlate the scores from different forms of tests. The invention of a test-scoring

machine in 1934 by IBM certainly made it easier for these two validation procedures to be implemented. And it was at this point in time that corporations began to get involved in test production and test scoring. The World Book Company had been publishing tests since the 1910's (Harcourt Brace and Company, 1994), but as other book publishers like Harcourt Brace Jovanovich and Houghton Mifflin became test publishers (Wardrop, 1976), this led to an increasing control that these companies would have not only on the development of tests but also on the curriculum that was taught in schools. Many teachers learned quickly that the test by which their students would be judged could be used and should be used (if students wanted to do well on the tests) as a guideline for what they taught in their classrooms. Through this trend, book publishers had their hand in the beginning and the end of curriculum development. That is to say that the same company that provided the test by which the students and the curriculum was evaluated published the textbooks which the teachers would use as a foundation of curriculum.

Lemann (2000) noted, "Walter Lippmann had predicted that if intelligence testing ever really caught on, the people in charge of it would 'occupy a position of power which no intellectual has held. . .'" (p. 69). ETS became that power. Other test publishers were not far behind. In 1947, the corporate involvement in testing became more prominent when Educational Testing Services (ETS) was founded (Dubois, 1970). ETS was formed from the Carnegie Foundation for the Advancement of Teaching, the College Entrance Examination Board, the Graduate Records Office of the Carnegie Foundation, and the Cooperative Test Service (Wardrop, 1976). Henry Chauncey sought to find a way to accurately compare students from across the country to one another so that they might have an equal opportunity at a college education. At that time, mainly the elite were given such opportunities to advance their

education. From the beginning, Chauncey and ETS desired that the SAT and the other standardized tests that they created would bring to life the vision that James Bryant Conant had of making education a way for anyone in America to have the opportunity to get a quality education and to develop their leadership ability (Calvin, 2000). Brigham had been one of the main obstacles in the creation of the agency, and his death in 1947 removed the final obstacle and opened the door (Lemann, 2000) for the project to move forward. Tanner and Tanner (1990) asserted, "A new era of standardized testing opened with the establishment of the Educational Testing Service in 1947, along with the uses and misuses of such tests for pupil sorting, tracking, and guidance" (p. 262). It was about this time that Arthur Bestor, a professor of history at the University of Illinois "proceeded to call for the use of standardized tests for pupil ability grouping and promotion. Such tests would also be used at the completion of compulsory schooling to dramatize that the continuance of one's education is not a right but a 'privilege bestowed upon the meritorious and the energetic'" (Tanner and Tanner, 1990, p. 337).

By the time that Harcourt Brace and World Book merged in 1960, World Book held the rights to publish the Stanford Achievement Tests, the Metropolitan Achievement Test, the Otis Mental Ability Test, and the Metropolitan Readiness Test (Harcourt, Brace, and Jovanovich, 1994).

But the growth of the testing industry was not credited to only the involvement of corporations. The expansion was also attributed to the social and political climate of the United States. According to Haney, Madaus, and Lyons (1993), there were five major social and political influences on the growth of educational testing: the launch of Sputnick in the 1950s, the Civil Rights Movement of the 1960s, the decline of SAT scores in the 1970s, the emergence of

the education reform movement of the 1980s, and national education reform proposals in the 1990s.

When the Soviets beat the United States to outer space, politicians blamed American schools. There was a push for more math and science courses and for student to be tested on just how much they learned in those classes. Then, throughout the late 1950s and 1960s, the Civil Rights Movement put the spotlight on education by highlighting the inequities that existed. Public Law 88-352, the Civil Rights Act of 1964, was made law on July 2, 1964 and outlawed segregation between Blacks and Whites in public places, including schools, in the United States. In order to ameliorate the situation, more and more students in schools were tested to be sure that the gap between schools for blacks and whites was narrowing. However the test scores from the SAT only proved that, no matter in which section of the country that the students were educated, black students from schools that were predominantly black did more poorly than blacks from integrated school (Lemann, 2000). Furthermore, it was during this time period that ETS first discovered that a variety of its tests were being used to keep blacks out of jobs. This did little to help improve the public relations nightmare though which ETS would soon live as a result of the declines of SAT scores in the 1970s. The perception of standardized testing and it impact on blacks and other minority students has been a subject of great concern throughout the last half century. Williams (1983) held that such minorities "tend to score low on standardized tests" and that "this is a barrier to their access to high quality education" (p. 198).

There were, however, detractors of this endeavor to classify and categorize students. One such noted critic was C. J. Karier. Wardrop (1976), in *Standardized Testing in the Schools: Uses and Roles*, cited that Karier called testing a way for a fourth branch of the government to exert its power and influence in American society. He was referring to corporate wealth which

insinuated itself into government by using standardized testing on a continuously more widespread basis. Perhaps Karier's stand was mischaracterized by Wardrop when Wardrop implied that American society's beliefs in democratic ideals made it prime for the standardized testing industry to flourish because it was seen as a "truly democratic way to 'sort out' individuals" (Wardrop, 1976, p. 12). Karier thought that tests "were biased in terms of social class, economic, cultural, and racial background. Their use in school served to block opportunity for the lower classes and immigrants. To the intelligence tests and to Lewis Terman, he [Karier] attributed responsibility for fashioning a system of tracking in the schools that reinforced social inequality" (Chapman, 1988, p. 8).

In 1983, *A Nation at Risk* recommended standardized testing as a means of school reform. The follow-up to that report, *High School: A Report on Secondary Education in America*, recommended that a Student Achievement and Advisement Test, similar to the testing system of the British, be implemented in Amercian schools (Haney, Madaus, and Lyons, 1993). It was because of such reports that the reform movement in education gained such momentum and has become the bane of the very existence of educators today. Neill and Medina professed that curriculum in schools today is influenced by standardized testing because the tests are the measuring stick by which teachers, administrators, schools, and school systems are judged (1989).

By 1994, testing was such big business that the mergers of several smaller companies such as The World Book, Psych Corp, and other smaller corporations had merged into one large company, Harcourt Brace (but called The Psychological Corporation), making it the "largest for-profit test publisher in the U. S. Only the nonprofit Educational Testing Service. . . rivals The Psychological Corporation as a test publisher" (Harcourt Brace and Company, 1994, p. 60).

Billions of dollars were at stake when policymakers and politicians began to talk of the measures necessary to hold educators accountable.

The Accountability Movement

Using assessment to hold teachers, schools, and states accountable is not new. According to Mazzeo (2001), in fact, during the late nineteenth century and early twentieth century, Kansas had state assessments in the eighth grade that were given to determine promotion to high school. The assessment was an essay examination that "were also used to allocate state educational resources fairly, shape teaching and learning in elementary schools, and reform—some say control—rural education" (Mazzeo, 2001, p. 375). Mazzeo (2001) pointed out one critical difference: there was no evidence that the scores were ever published in the newspaper for everyone to see how different schools compared to each other or whether or not the schools needed improvement. During the 1910's and 1920s, accountability testing took the form of teacher evaluation and state supervision in five states. But these types of testing policies were not continued. The push for accountability that still pervades educational reform today took root in the 1970s and 1980s. From 1967 to 1973, the number of states that had testing policies rose from only four to thirty-three (Mazzeo, 2001).

At this time, Mazzeo (2001) asserted, the culture in the United States is one of "'no excuses'" (p. 390). Kliebard (2002) emphasized, "Policy makers continue to try to improve school practice, of course, but the most widely touted reform takes the form of specifying rigorous achievement standards accompanied by high-stakes testing. When students do not measure up, school officials are urged to deny them promotion or graduation. Presumably positive results will ensue if children and youth are so coerced, but the actual outcome of such a policy is not clear" (p. 1). Kliebard (2002) went on to assert that by itself testing is not reform at

all. "Frequently, schools, school districts, and state departments of education seek to achieve excellence simply by testing alone, by raising minimum requirements on such tests, or by simplistic mechanisms such as increasing graduation requirements. Because support systems for students are lacking or inadequate in some of these cases, excellence is not actually advanced, only proclaimed" (Kliebard, 2002, p. 120). Kliebard (2002) maintained, "Recent emphasis on high-stakes testing may also serve the purpose of keeping teachers in line" (p. 129).

As is evident by such legislation as No Child Left Behind, the current trend in society is to hold students, teachers, schools, and school systems accountable. No Child Left Behind requires that states enact a policy of assessment in grades 3, 5, and 8 each year in order to track progress and determine Adequate Yearly Progress (AYP). According to Bourque (2004), this area of the law is very "unclear and ambiguous" (p. 233). Even though proponents of the law give the perception to the public that there are absolutes, and no excuses for states, there is a great deal of room for maneuvering on the part of the states when it comes to determining AYP. There is an absolute goal that must be reached by the 2013-2014 school year, but states can set their own pace in getting there beginning with the baseline data that they collected during the 2001-2002 school year. States can choose to set gradual, steady goals that reach the ultimate target of one hundred percent passing rate, or states may choose to make goals that require little percentage improvements in the first several years and require larger gains in the final years leading up to 2013-2014. Annual report cards that reflect how subgroups of students perform on the individual state's assessments must be published, according to the law.

One more loophole is that the subgroup must have at least forty students in the school in the grades that are being considered before the subgroup's scores count for determination of AYP. For example, in a rural elementary school that has only 450 students, if there are fewer

than 40 students with disabilities in grades 3, 4, and 5, then the progress of those students is not considered for AYP determination. Essentially, the law says that is does not matter if those individual students progress or not. They do not comprise a group that has statistical relevance. Therefore, a school can not provide a quality, meaningful education to their students with disabilities and still make AYP as long as there are not forty or more such students in grades 3, 4, and 5 in that school. Yet the politicians can still project the image to the people of that rural community that they are in control and are making a difference in student achievement because the law exists.

Summary

This chapter began with a look at the first examinations that were given in the United States, the Civil Service Examinations, and how the idea of using tests to determine vocation and educational course evolved in this country. The chapter further investigated the influence of Europeans such as Alfred Binet on the psychological testing movement in the United States. Along with examining mental tests and their uses in the early part of the twentieth century, this chapter also presented the arguments on both sides of the debate over testing: is it appropriate to categorize students using a test score or not. There were points of view of both proponents and detractors of testing presented in this chapter. Finally, the evolution of testing from a means to guide individuals to the money-making venture that it became in the middle of the century to the accountability tool that it has become in recent years was probed.

Chapter 3

SEVEN STANDARDIZED TESTS

This chapter will provide the background of seven major assessments: the Stanford Achievement Test, the Scholastic Achievement Test, the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the California Achievement Test, the Comprehensive Tests of Basic Skills, and the National Assessment for Educational Progress. When information was available, a description is provided for each assessment regarding the date of original administration, the dates of publications of revisions and various forms, and the changes represented by the revisions of the assessments. Such detailed information was not available for all seven assessments. After many efforts to obtain primary source documents of actual assessments and supplementary publications such as technical manuals or guides, secondary sources were used in some instances.

The Stanford Achievement Test

The first edition of the Stanford Achievement Test was published in 1922. While Harcourt Assessment stated on its website that 1923 was the first year of publication (2006), a first edition of the assessment dated 1922 and published by World Book Company was located in the Main Library at the University of Georgia. This first edition of the assessment included Forms A and B for grades 2 through 8 (Bryan, 1965). The subtests included reading, spelling, arithmetic, nature study, science, history and literature (Bryan, 1965). According to Bryan (1965), some parts of the test were multiple choice while others required students to write responses. The norms that were provided in the 1922/23 version were revised in 1925 with a much higher number of cases, and a "revised Manual of Directions was printed" (Bryan, 1965, p.

111). With the exception of those additions, no further changes were made until 1929. "Revisions were published in 1929, 1940, 1953, 1964, 1973, 1982, and 1989" (Harcourt Brace Educational Measurement, 1997, p. 7).

From 1929 to 1931, the revisions that were noteworthy included expansion to ninth grade ability and some structural changes to the test, like the division of the literature and history test into two separate tests and the printing of two columns per page in the test booklet (Bryan, 1965). In the 1930s, The Guide for Interpreting was also added to the administrator's manual. In 1940, though, the biggest reform in the test took place: there were five forms of the assessment introduced with approximately eighty percent new items (Bryan, 1965). The Primary, Intermediate, and Advanced Batteries were the levels that were offered (Bryan, 1965). "The most impressive feature of the 1940 revision was the norming program involving the testing of approximately 300,000 pupils in 173 communities in 32 states, from which a random sample of 50, 955 cases was drawn" (Bryan, 1965, p. 113). Interestingly, unlike other editions of the administrator's manual, the Guide for Interpreting that was promised for the 1940 edition of the assessment was never published (Bryan, 1965).

By the time the 1953 version of the test was published, "there were four almost entirely new batteries: the Primary Battery for grades 1.9 to 3.5; the Elementary Battery for grades 3 and 4; the Intermediate Battery for grades 5 and 6; and the Advanced Battery for grades 7, 8, and 9" (Bryan, 1965, p. 1130. Forms J, K, L, M, and N were offered. The 1964 edition of the assessment was the fourth extensively revised edition of the assessment. It was "the product of five years of research and developmental work" (Bryan, 1965, p. 115). Four forms of the test were published (W, X, Y, and Z), and the organization into five batteries provided a "better at-grade coverage of content and skills" (Bryan, 1965, p. 116). Most of the items were new, and

the metamorphosis from the very subjective test in its first edition to a virtually completely objective test was complete (Bryan, 1965).

It was in 1973 that the sixth edition of the Stanford Achievement Tests was published. Because the first edition of the assessment took place in 1923, the 1973 version was nicknamed "'The Golden Anniversary Edition'" (Passow, 1978, p. 102). The number of levels had grown from the original two to six different levels in 1973. In the seventh edition in 1982, an option writing portion was added, but that portion was not included in the eighth edition in 1990 (Brown, 1992).

According to the Harcourt Assessment website, the ninth edition of the Stanford Achievement Test was published in 1996 as a result of "(a)significant changes that have occurred in the school curriculum, (b) the need to update norms and interpretive materials, and (c) the need to provide for the continuous assessment of achievement in the major skill areas" (Berk, 1998, p. 925). The tenth edition was first published in 2003. The publishing company cited the same reasons for the revision as it did for the publication of the ninth edition (Carney, 2005). The assessment continues to be published by Harcourt Assessments (2006).

The Scholastic Aptitude Test

Donlon (1984) stressed that, at the end of the nineteenth century and at the beginning of the twentieth century, educators, particularly those in secondary schools, were faced with the difficulty of preparing students for college. There was such diversity in the curricula of secondary schools across America that it made it nearly impossible for administrators of those schools to prepare their college-bound students for colleges with equally diverse degree programs and admissions requirements. Thus, testing was a way to save time, money and effort for college admissions officers, secondary school administrators, and students. Out of this need

for an easier method of getting students into colleges and universities, the College Entrance Examination Board was born. It was formed at a meeting of the Association of Colleges and Preparatory Schools of the Middle States and Maryland in Trenton, New Jersey, on December 2, 1899.

Even though the primary focus of the College Board was not to create tests (rather, they strove for better communication between secondary schools and colleges), testing soon assumed the position as the primary responsibility of the organization, In June of 1901, 973 students took the first College Board exam at 69 test centers across the United States. On June 23, 1926, the first Scholastic Aptitude Test was given to 8,040 students in nine subtests: definition, math problems, classification, artificial language, anotnyms, number series, analogies, logical inference, and paragraph reading (Donlon, 1984). The test originally consisted of both multiple choice and essay questions. In 1929, the scores on the subtests were divided into a math score and a verbal score, the format with which most people are familiar today. The test was primarily written by Carl Brigham from 1926 through 1941. "In Brigham the College Board found a man whose credentials, both personal and professional, were ideal for the task not only of designing a test useful to admissions officers but of bringing skeptical headmasters and college presidents around to the idea of accepting the test" (Valentine, 1987, p. 34).

During the 1930s the SAT met with hardship; test participation dropped by 35 percent in a period of five years from 1931-1936. However, this did not change the opinion that the College Board held a very high position of influence. In fact in a Report to the Executive Committee of the College Entrance Examination Board, Carl Brigham (1933) asserted, "The Board occupies its present position due to its success as a political and governmental institution" (p. 1). Even the members of the Board itself understood the power that the Board held at a time

in history when test participation was on the decline. And this decline did not dissuade the College Board from continuing the test, and in April 1941, the students who took the SAT became the norm group for all future forms of the exam until a "recentering" of the test in 1995 (Donlon, 1984; ETS, 2007).

Even the administrations of the SAT underwent tremendous change during this time period. There was an increase of participants in the 1940s, and it became necessary to give an April administration (designed for scholarship applicants only) and a June administration (designed for non-scholarship applicants). But it was in 1942, because of World War II, that Harvard, Yale, and Princeton started the college year in June or July requiring students to take a multiple-choice only version of the test in April, thereby eliminating the different administrations that separated the students by socioeconomic status. College admissions officers decided that the multiple-choice items were sufficient, and the June essay questions were eliminated after 41 years of use (Donlon, 1984).

Although the SAT was the primary test that the College Board administered, World War II brought a new field of opportunities to the organization. The Board was asked to design tests for officer candidate training (V-12 Testing Program) and then to design tests that would be suitable for armed forces veterans who would want to enter college after their return from World War II. It was then that the Board began to experiment with creating tests for scholarships for corporations such as Westinghouse and Pepsi Cola and for the Military Academy, the Naval Academy, and other governmental institutions. However, the Board was happy to turn this over to ETS in 1947 (Donlon, 1984). In 1948, the administrative tasks regarding the compilation of the assessment fell under the jurisdiction of the Educational Testing Service (Donlon and Angoff, 1984).

In the 1950s, 1960s, and 1970s, computer technology influenced the number of tests that could be administered and the way that test scores were reported (Donlon, 1984). Prior to the 1940s, there was only one administration of the SAT per year; however, by the late 1950s there were three per year. And by 1977, there were six each year—twelve if the Sunday administrations were counted. In June 1980, the College Board began to offer Sunday administrations for those who could not take the test on Saturday due to religious reasons (Donlon, 1984).

In 1971, some of the tests with which people are the most familiar today came into existence. For example, the Preliminary Scholastic Aptitude Test (PSAT) was adopted by the National Merit Scholar Program and was renamed the PSAT/NMSQT. It was used as a qualifying exam for the National Merit Scholar Program. By the late 1970s, the PSAT, Advanced Placement tests, College Placement Tests Program, Comparative Guidance and Placement Program, and the College Level Exam Program (CLEP), and the Test of English as a Foreign Language (TOEFL) were all in place (Donlon, 1984). The College Board had broadened the scope of the services that it could offer. Once can tell from the kinds of tests that the College Board began to offer that it was responding the growing needs of the kinds of students who were going to college. No longer were there only white males attending colleges and universities. College and university campuses were becoming places of great diversity.

The great diversity on those campuses dictated that the kinds of questions that were on the SAT should change. The College Board responded by changing the questions to reflect the increasingly diverse colleges that relied on the assessment (Donlon, 1984). Since the 1970s, there had been substantial changes in the SAT. In 1971, the College Board published a technical report written by Angoff and Dyer. Angoff and Dyer (1971) claimed that the SAT was "in an

unusually strategic position to exert a significant influence on American secondary school education" (p. 6). Angoff and Dyer (1971) asserted that this was the case because schools based their curriculum on what they anticipated as the content of the upcoming forms of the SAT. According to Angoff and Dyer's theory, then, the changes in the SAT should have resulted in a great shift in the curriculum of America's schools. Since the last few decades have seen the rise of multicultural aspects of the curriculum in America's schools, specifically history and literature, then Angoff and Dyer's theory would have dictated that the SAT contain questions that reflected such a shift in these curricula. However, there are ample critics who maintained that the SAT certainly still has not resolved its issues with minority test takers in terms of questions that are culturally sensitive. Instead, the test is still considered to be largely biased against minorities.

The Iowa Test of Basic Skills

In November 1928, the Iowa Test of Basic Skills got its start as an academic competition in the state of Iowa (Peterson, 1983). The state had plenty of opportunities for athletic teams to compete, but there was a desire to provide a chance for students to compete with each other. According to Peterson (1983), "The prime objectives of the program were: *first*, the improvement of educational measurement, and *second*, the stimulation of scholarship, in Iowa public secondary schools" (p. 2). During the first year, 223 schools participated and 40,000 students were tested, and in the second year, 360 schools participated. By the third year, 395 schools participated (Peterson, 1983). The competition provided that from each school, the two students with the highest raw score points in each subject were able to participate in a district competition. Then, the winners from the district competitions traveled to Iowa City to participate in the next level of competition (Peterson, 1983).

After the first two years, the district competitions were eliminated, and the contest was renamed the State Scholarship Contest. The final contest was still held in Iowa City, and one of the prizes was a jeweled gold key, studded with a ruby for first place and a pearl for second place. "These emblems, especially the jeweled keys, were highly coveted and treasured. Doubtless many may still be found in old family jewelry boxes" (Peterson, 1983, p. 6). The tests were then renamed the Iowa Every Pupil Test.

E. F. Lindquist, the creator of the program, asserted that the tests were designed with the high schools of Iowa in mind and that the norms were based on a higher number of students that the average standardized test. As a result, he claimed, "'They may be used for all of the purposes for which the usual standardized test might be employed, and because of these superiorities they can be so used with much greater effectiveness'" (quoted in Peterson, 1983, p. 4). From 1932-35, "A noncompetitive basis of participation provided an alternative for administrators who wanted the values of annual measurement but felt that competition had fulfilled initial purposes" (Peterson, 1983, p. 9). It was at that time that "[s]chool personnel needed and wanted detailed interpretative materials and suggestions for using the test results" (Peterson, 1983, p. 9).

Lindquist (1944) made note of the fact that the Iowa tests were economical because the test booklets had all nine subtests in one booklet, and the tests were loaned to the schools that use them. Another economical benefit was the tryout of new test items. "In the Iowa program we plan to overcome this difficulty by requiring every school to devote fifteen or twenty minutes in each annual program to the tryout of new materials for the subsequent programs" (p. 91). In the first years of the assessment, norms for the ITBS were established yearly so that they were always up to date (Lindquist, 1944). According to Peterson (1983), "[i]n this and all later phases

of the Iowa Testing Programs, precisely formulated and executed procedures have been fundamental to the success of the undertaking" (p. 12). The Iowa Testing Programs published an intelligence test in 1934, but it only lasted for one year. It was not clear why it was not continued, but the expense and the need to concentrate on the publication of a new elementary battery in 1935 were probably reasons that impacted this decision (Peterson, 1983).

The Iowa Testing Program became popular outside the state, and it was decided that neighboring states could participate on a noncompetitive basis beginning in 1935 (Peterson, 1983). "In addition, arrangements were made with the Bureau of Educational Research and Service to sell the Every-pupil Tests nationwide for independent administration at any time of the year" (Peterson, 1983, p. 23). The test writers (of whom E. F. Lindquist was the primary), designed test materials that would aid in educational guidance and individualized instruction (Peterson, 1983). "It was promised that 'the results of tests provided in succeeding years will be made *comparable* to those of the tests used this year, and high comparability from test to test will characterize each year's battery. It will thus be possible to keep a cumulative record of the *progressive* development over several years' time of the skills measure for each pupil, and hence to base educational guidance upon a much more meaningful and reliable description of the pupil than could possibly be secured from any single test or battery of tests'" (Peterson, 1983, p. 30).

According to Peterson (1983), in 1940, Form L was released, and during the next eleven years, Forms L through T were released, one each year, and the test expanded to include grades three through five as well as six through eight. It was also around this time (in 1949) that the directorship of the program shifted from E. F. Lindquist to A. N. Hieronymus. In Hieronymus' first year, the Iowa State Department of Public Instruction asked to give the ITBS to all Iowa pupils in grades 6-8 "for the purpose of comparing achievement in the one-room rural schools

with that in graded schools in communities below 10,000 population" (Peterson, 1983, p. 50). Because of the confidence that Lindquist had in the quality of his ITBS, he anticipated that the out-of-state demand for the test would necessitate the use of a publishing company to accommodate "commercial distribution by a well-established and highly respected publisher of education materials. This he arranged in 1940 with the Houghton Mifflin Company of Boston" (Peterson, 1983, pp. 50-51). This partnership allowed for the test to be used widely in states other than Iowa. Missouri was the first state outside Iowa to utilize the ITBA. Iowa schools still received the assessment in the same format as before, but "out-of-state schools in general bought the test materials at a higher price directly from Houghton Mifflin and received no reporting or statistical services" (Peterson, 1983, p. 50). The lack of scoring services kept the assessment largely affordable for most school districts while removing the burden of publishing such a high volume of assessments from the Iowa Testing Programs.

In 1948, "Houghton Mifflin issued a machine-scorable answer sheet for Tests A, B, and C of ITBS Forms L, M, N, and for Test D of Forms O and P" (Peterson, 1983, p. 52). Science Research Associates of Chicago contracted with Houghton Mifflin to offer scoring services for Forms R and S and this continued through 1956 (Peterson, 1983). This addition made the results of the assessment more useful for test users. Having scores returned in a quick manner with score reports that gave practical information about the test taker made the assessment attractive to school districts.

According to Peterson (1983) in September 1942, the Fall Testing Program and the ITED (Iowa Tests of Educational Development) were initiated. The guidelines are as follows (p. 55):

Important criteria of an effective wide-scale program of testing for evaluation, guidance, and the individualization of instruction.

1. The tests used should measure as directly as possible the attainment of the ultimate objectives of the entire school program.
2. All of the tests should be administered, under standard conditions, to the entire student body. . . .
3. The program must provide for the measurement of growth. . . .
4. The tests used should measure the more permanent of the changes produced in the pupils. . . .
5. The test results should not be usable in the rating of individual teachers.
6. The description of the pupils' educational development provided by the tests must be expressed in readily interpretable form. . . .
7. The test profile for each individual pupil must be readily available at all times to each of his teachers and counselors. . . .
8. The measure derived must be highly comparable from test to test. . . .
9. Each of the tests used must yield highly reliable measures of the abilities of the individual pupil. . . .
10. The testing program ideally should impose no clerical or statistical burdens whatever upon teachers and administrators, and in all other respects should involve the minimum of administrative inconvenience. . . The total cost of the services of the program must be within the reach of the majority of the public schools. (p. 55).

These guidelines provided for the standardization of the assessment that kept the results valid and reliable for measuring student progress over time.

Peterson (1983) also emphasized that test security became the highest concern. In order to promote greater security, the mission to have a scoring machine at the University of Iowa began in September 1952. The machine was put into use in March 1955. It cost $200,000 to build. "It made possible, within a matter of months, large-volume test processing service to schools across the nation, on a variety of test batteries and at a reasonable charge" (p. 52).

In 1950 and 1951 Forms S and T were released. Form Q was reused in 1952, Form R in 1953, and Form S in 1953. In 1955, the ITBS issued the Multilevel Edition that incorporated six different grade levels. "'We would start each grade at the point in the test when the items began reflecting the objectives of that grade at an appropriate level of difficulty. The test for a given grade would stop at a point in the test when the items were no longer appropriate in content or in difficulty'" (Peterson, 1983, p. 130). Hieronymus sought to develop a test that would be for a single grade level, but there were limits in reality for the test to remain efficient and for an agency to have the resources necessary to create such a test (Peterson, 1983). In 1970, however, grade designations were removed from the tests, but schools could choose levels to give to students (Peterson, 1983). In 1972, the Primary Battery was introduced, and in 1979-1980, the Early Primary Battery for Kindergarten and first grade was released (Peterson, 1983).

According to the website of Riverside Publishing, the Iowa Test of Basic Skills is now published by Riverside Publishing, a subsidiary of Houghton Mifflin Company. Houghton Mifflin has been involved in the standardized testing industry since World War I when it published the Stanford-Binet Intelligence Scale. The company worked closely with E. F. Lindquist from the beginnings of the Iowa Testing Program's branching out to all states (2004).

The Metropolitan Achievement Test

According to the history of Harcourt Assessments published on the company's website, the first edition of the Metropolitan Achievement Test (MAT) was published by World Book in 1932. However, Findley (1953) claimed that the first edition of the assessment was in 1931. And Linn (1985) maintained that the first edition was published in 1937. The test was originally published "to assess the achievement of students in New York, NY City public schools" (Wolf, 1978, p. 67). But some scholars have been complimentary of the contributions that the battery had made to the field of educational measurement over the years. The battery was responsible for an "expansion of the test manual as an aid to users, the thorough and expert standardization of the tests, the provision of varied norms, including some for special types of schools, emphasis on the desirability of using tests early in the year for instructional purposes, inclusion of a section on elementary science" (Findley, 1953, p. 48). In a subsequent review of the MAT, Findley (1965) also found that the 1959-62 series of the MAT was not entirely different from the earlier editions with the exception that the High School Battery was added. Findley (1965) hailed the scope of the assessment as well as its "measurement of important outcomes" as much improved from earlier versions (p. 67). The biggest improvement that Findley touted was the user's manual (1965).

In 1978, the fifth edition of the test was published (Linn, 1985). Linn (1985) asserted that the "most significant change from earlier editions was the introduction of the 'two component system'" (p. 965). This system simply meant that the batteries of the MAT included both an instructional set of tests and a survey set of tests. The seventh edition of the MAT, though, was a major overhaul of the MAT6, according to Finley (1995). Finley (1995) stated that the publisher cited "changes in school curriculum, changes in assessment trends and

methods, and the need for updated testing materials, normative information, and interpretive materials" (p. 603).

Through a series of corporate acquisitions, World Book became Harcourt Brace and World and eventually simply Harcourt Assessments. In 2000, the eighth edition of the MAT was released by Harcourt Assessment (2006). The publishers cited that the changes in the eighth version of the test were "undertaken in order to: (a) align the content of the test with the most recent curricula, (b) expand the scope of the test to include items designed to measure both basic and higher order thinking skills, and (c) update the norms" (Lukin, 2005, p. 7).

The California Achievement Tests

The California Achievement Test was originally known as the Progressive Achievement Tests (Carpenter, 2007) and dated back to 1933 under that title (Womer, 1978). According to Merwin (1965), however, the first version of the Progressive Achievement Test was published in 1934. "The tests measured three basic subject matter skills—reading, arithmetic, and language" (California Testing Bureau, 1957, p. 5). In 1937, the first revision of the assessment appeared (California Testing Bureau, 1957). Again in 1943, the test was published as the Progressive Achievement Test (1978). It was with this edition of the PAT that percentile norms were first published (California Testing Bureau, 1957). The California Achievement Tests (forms AA, BB, CC, and DD) were first given in 1950 and were made up of multiple choice assessments that measure student development in Reading, Spelling, Language, Mathematics, Study Skills, Science, and Social Studies. The 1950 version of the test that first bore the name California Achievement Test differed from its previous incarnations in that there were both content and structural changes. The structural changes consisted of "splitting of the primary level battery of the 1950 tests into two batteries called Upper Primary and Lower Primary" (Merwin, 1965, p.

17). In terms of content, the changes were in the reading and mechanics portions of the tests (Merwin, 1965).

The next versions of the assessment were released in 1957 (the WXYZ Series), in 1971, in 1977 (Forms C and D), and in 1985 (Forms E and F). Ernest W. Tiegs and Willis W. Clark were the authors of the assessment at the time of the 1957 publication (California Test Bureau, 1957). The 1957 Technical Report maintained that "one of the main purposes of the 1957 tests is to continue the aim of earlier work on this series 'to develop a battery of diagnostic tests rather than another survey-type batter yielding but single subject area scores'" (Merwin, 1965, p. 17). The 1971 edition of the assessment was a complete overhaul from the 1957 version. "Changes were made in every test at every level. Some of these were minor; but some were extensive. . . ." (Bryan, 1978, p. 35). Forms C and D of the assessment were touted as "comprehensive [which] means pre-reading, reading, spelling, language, mathematics, and reference skills" were tested (Rogers, 1985, p. 243). Another major change in this version of the assessment was that they were marketed as both a norm-referenced and a criterion-referenced test. Reviewer Willson (1985) maintained that it was developed in the "classical framework" of a norm-referenced test (p. 247).

A performance assessment component also became available beginning with the fifth edition of the assessment which was published in 1993. This assessment consisted of constructed response items in Language Arts, Mathematics, Science, and Social Studies (Nitko, 2004). The sixth edition of the assessment is also called the *TerraNova*, Second Edition, Forms C and D, and it was released in 2001 (Spies and Plake, 2005).

According to the publisher CTB/McGraw-Hill, the tests give details about how students compare to others in the norm group and about the individual student's instructional needs. The

assessment is not based on a specific curriculum and "is intended to measure a student's understanding of broad concepts" (CTB/McGraw-Hill, p. 2-1).

The Comprehensive Tests of Basic Skills

According to the Communications Department of CTB McGraw-Hill (2007), the first forms of the Comprehensive Test of Basic Skills (Forms Q and R) were given in 1969. The assessment was marketed "with a variety of answering formats at each level" (Findley, 1978, p. 41) from its first edition. Forms S and T followed with versions issued in 1973 and 1975. This expanded edition of 1975 was "expanded downward" (Findlay, 1978, p. 41) so that grades two through twelve were included in the batteries. Additionally, science and social studies tests were added (Nitko, 1978). Nitko (1978) also cited that there was an effort to reduce racial and ethnic bias in this expanded edition.

It was in 1981 and 1982, respectively, that Forms U and V hit the market as the third edition of the assessement (Linn, 1985). In 1989, the 4th Edition of the CTBS was marketed. With CTB McGraw-Hill as the publisher of the CTBS as well as the California Achievement Test, the 5th edition of this assessment, too, became known as *TerraNova* Form A when it was released in 1997 (Monsaas, 2001). *TerraNova* Form B, also known as a 5th Edition of the CTBS, was made available in 1998. By the time that the *TerraNova* Forms C and D were published as a parallel to the CTBS, it was also the sixth edition of the California Achievement Test (Spies and Blake, 2005). Finally, in 2007, the name Comprehensive Test of Basic Skills was lost altogether when the test was called simply *TerraNova* Form E.

The National Assessment for Educational Progress

The National Assessment for Educational Progress is also referred to as "the Nation's Report Card" and "is the only nationally representative and continuing assessment of what

Amercia's students know and can do in various subject areas" (National Center for Educational Statistics, June 2006). According to the *NAEP Overview* from the National Center for Educational Statistics, the specifics for the assessment and the frameworks were set by the National Assessment Governing Board which is not affiliated with the United States Department of Education but is appointed by the Secretary of Education (June 2006). The assessment at the national level includes students from grades 4, 8, and 12 and includes questions that reflect the latest developments in the field of assessment (National Center for Educational Statistics, June 2006). Constructed-response questions and questions from the arts and sciences are used to determine a students' proficiency in hands-on tasks (National Center for Educational Statistics, June 2006).

The National Assessment for Educational Progress had its beginnings in a request of United States Commissioner of Education Francis Keppel to Ralph Tyler about assisting in determining a method for evaluating education in the United States (Jones and Olkin, 2004). From 1963 to 1966, a series of conferences were held that included the Carnegie Corporation, which contributed $100,000 to the Exploratory Committee for the Assessment of Progress in Education (Jones and Olkin, 2004). The Carnegie Corporation and the Ford Foundation earmarked another $2.8 million for the project, and it wasn't until 1966 that the United States Office of Education contributed any money to the cause--$50,000 for additional conferences to be conducted by the University of Minnesota (Lyle and Olkin, 2004). According to the National Center for Education Statistics, the first administration of the National Assessment for Educational Progress took place in 1969. "The launching of the National Assessment for Educational Progress (NAEP) occurred in 1969 only after political safeguards were built into the assessment process guaranteeing that it would not become a national test linked to national

standards with invidious state comparisons under the control of the federal government" (Epstein, 1996, p. 22). The assessments that began in 1969 stayed at the national level and the subjects included citizenship, science, and writing in the first year. Beginning with the second year of assessments, literature and reading were added, and music, social studies, writing, and mathematics were added in the following years (National Center for Education Statistics, August 2006). From the beginning, great pains were taken to be sure that states could not be compared or that state average scores would be released (Robinson and Brandon, 1994).

But those guarantees to prevent the assessment from becoming a way to compare states did not last. Epstein (1996, p. 24) claimed

> It is out of the tensions arising from the politics to control the policies and direction of NAEP that the original guarantees made to educators were breached. Consequently, NAEP has been turned away from its original intent of providing reliable information for the purpose of informing educational policy toward becoming a more politically potent assessment instrument intended to shape and influence the direction of educational policy in the nation.

Beginning in 1969, the assessment was administered by the Education Commission of the States. In 1978, Congress had passed Public Law 95-561 to authorize NAEP (Epstein, 1996). In this law, Congress funded NAEP "'by grant or cooperative agreement with a nonprofit education organization.' Excluding a procurement contract as an option implied a recognition by the Congress of the need to fund a National Assessment with minimal control from the federal government" (Epstein, 1996, p. 26). In 1983 when the Educational Testing Service (ETS) won the grant for administration, funding for the NAEP shifted from a federal grant to a federal

contract and this "signified an increase in federal control over the direction and operation of the project" (Epstein, 1996, p. 25). ETS, winning the contract, was able to make the National Institute of Education's desire that NAEP have more influence over educational policy come true (Epstein, 1996). "As the most powerful testing service in the nation, ETS understood the value of introducing competition to increase the importance of a test and by expanding NAEP's influence, ETS continued to expand its influence in educational testing. Therefore, it is not surprising that ETS embarked on a course to raise the stakes of the assessment through state comparisons" (Epstein, 1996, p. 29).

It was to the economic and political benefit of ETS to have administrative control over the NAEP. Although ETS was a non-profit organization, the revenue that were generated by the various assessments provided by the agency acted to provide ETS with money to conduct research in a variety of areas. As for the political influence, the more that ETS was in control of the major testing ventures in the United States, the more influence it could exert over curriculum and policy. "ETS advanced its agenda for state NAEPs in a climate of renewed energy for state-based educational reform which had been sparked by the release of national reports such as *A Nation at Risk* and *Action for Excellence*" (Epstein, 1996, p. 29). Over the years, ETS subcontracted various services to other companies such as Westat, Incorporated, American Institutes for Research, Pearson, and National Computer Systems (National Center for Education Statistics, 2003). And in 1985, when the NAEP was transferred from the National Institute of Education to the National Center for Education Statistics, "ETS had demonstrated that a state NAEP was feasible and all that was needed was a change in the legislation authorizing NAEP to make it possible" (Epstein, 1996, p. 30).

In 1986, the National Governor's Association, which was then chaired by the Governor of Tennessee Lamar Alexander, "endorsed the collection of state level achievement data as part of their plan for educational reform" (Epstein, 1996, p. 30). According to Mosher (2004), the 1988 reauthorization of NAEP saw Congress create the National Assessment Governing Board. This move was at least partially in response to the desire of many to report the results of NAEP "in terms of the proportions of each group who exceeded one of three 'achievement levels' (basic, proficient, or advanced)" (p. 331). Mosher (2004) also points out that there were those "who wished the schools to be seen as doing badly" (p. 332). NAEP was seen as even more credible because the political climate was such that bad news was welcomed by many (Mosher, 2004). "By pushing for the collection of state comparative data, ETS had been pivotal in charting an altered course for the National Assessment that educators had been guaranteed would never happen, a course that was remarkably similar to predications the educators had warned against in the 1960s" (Epstein, 1996, p. 30).

According to Epstein (1996), in the late 1980s when the new legislation for NAEP was under consideration by Congress, the American Association of School Administrators (AASA) and the National Parent Teacher Association were "opposed to raising the stakes of NAEP through state comparisons when the literature and evidence was clear about the effects high stakes tests have on influencing policy and curriculum" (p. 31). But in 1988, Congress passed legislation that, for the first time, allowed NAEP to collect data from the states that would hold them accountable for student achievement—"NAEP was being turned into an instrument of accountability" (Epstein, 1996, p. 34).

One of the amendments to this 1988 legislation, the Hawkins-Stafford Amendment, gave the National Assessment Governing Board [NAGB] the responsibility of "'identifying

appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment' (P. L. 100-297). Apparently, these twenty words gave this independent board statutory authority to set national standards which incidentally nullified another of the safeguards that had been originally built into NAEP, a guarantee that the National Assessment would never be used an an instrument to set national standards" (Epstein, 1996, p. 36).

Epstein (1996) argued that then President George H. W. Bush, along with all 50 governors (including Governor Bill Clinton of Arkansas), in 1989, determined that is was appropriate to set national educational goals. President H. W. Bush called his plan American 2000, and the subsequent Clinton administration renamed is Goals 2000. Goal 3 of this plan called for students leaving grades 4, 8, and 12 to demonstrate proficiency in subjects such as English, mathematics, foreign languages, science, and economics. Since the NAGB was given the responsibility to identify "appropriate achievement goals" by law, they set three achievement levels: basic, proficient, and advanced (Epstein, 1996). "Establishing these cutoff scores caused a firestorm of controversy between NCES and NAGB over the methodology used to set the cut points" (Epstein, 1996, p. 27). The National Education Goals Panel issued a disclaimer that "'The NAEP data should be interpreted with caution. . . .[T]he methods used to derive the NAGB achievement "cut points" (i.e., the points distinguishing the percentage of students scoring at the different achievement levels) have been questioned and are still under review' (NEGP, 1994, p. 134). However, the disclaimer does not negate the fact that these standards are being used and published to measure progress toward meeting the national education Goal 3" (Epstein, 1996, p. 37).

No Child Left Behind has very specific mandates regarding the states' administration of the NAEP. Beginning in the 2002-2003 school year, the assessments were required every two

years in grades 4 and 8 in order for states and local education agencies to continue to receive Title I funds (National Center for Education Statistics, August 2005).

Summary

This chapter chronicled the background of seven of the most popular standardized tests of the twentieth century. From the very earliest version of the Stanford Achievement Test to the relative newcomer, the National Assessment for Educational Progress, the examinations discussed here have shaped education in the United States. In many cases, the current versions of many of these assessments are different in form and in function from the initial authors' intent, but they have all withstood the criticisms and changes to become some of the most valid assessment instruments available in the United States.

Chapter 4

THE APPROPRIATE AND INAPPROPRIATE USES OF STANDARDIZED TEST SCORES: SCHOLARS OF EDUCATIONAL MEASUREMENT

Cannell (1989, p. 39) asserted

> The norm-referenced tests now used to assess public educators were never designed for such a task. They have evolved from being instructional and curricular aids into instruments of public accountability because of external political 'accountability' pressures.

This chapter will focus on the literature that has been published in the field of educational measurement. In this literature, the purposes of tests and the use of their test results that were considered to be appropriate will be examined alongside those uses and purposes that were deemed unsuitable. The work of leading test experts in the field such as John Dewey, Lewis Terman, E. F. Lindquist, George Madaus, and James Popham will be analyzed. Both the positive view of testing as well as the harmful effects of testing and the use of test scores will be discussed. The leading test designers often had opposing views as to the appropriate use of tests and their results. This chapter will explore the differing opinions as well as chronicle the scholars who had changed their minds regarding the role that testing should play during the course of their lives and study within the field of educational measurement.

From early in the twentieth century, leading psychometricians asserted what they believed to be the appropriate and inappropriate uses of test scores. Pressey and Pressey (1922) advocated for using tests in conjunction with other pieces of information about a student to make such important decisions such as those regarding promotion and retention or for determining the

competency of a teacher or the quality of a school. They stated, "Tests are not infallible. . . . And they must be used not blindly, but sensibly and intelligently, with due consideration for other sources of information" (Pressey and Pressey, 1922, p. 70). Pressey and Pressey (1922) even held that tests were a good way to measure student ability and that a teacher should use tests in combination with a student's everyday performance to determine the strengths and weaknesses of the student. They asserted that an appropriate use of test scores was to test the effectiveness of new teaching methods. Further, Pressey and Pressey (1922) asserted that tests were a good way for supervisors to determine the progress of students. However, they were also adamant that the teacher's employment should never be based on the test's report of student progress. Pressey and Pressey (1922) maintained that teachers should never be fired because their students do not perform well on tests. Tests had a bad reputation among teachers because they were being used in this way. They stated, "if tests are to be used only as evidence against the teachers they had best not be used at all" (p. 30).

Lewis Terman's philosophy seemed to be consistent with this viewpoint. Terman (1923) contended, "There is no warrant for grading all pupils rigidly on the basis of mental age, even if mental age is the most important single factor. A pupil's fitness for a given grade depends in some degree upon his previous instruction, his health, his physical maturity, his industry, and his attitude toward school work" (p. 11). Terman called for the test score to "be taken as the point of departure for further study of the pupil" (Terman, 1923, p. 25). For example, in the event that a student's test score did not match with what the teacher knew of the other attributes of the student, then more study (and perhaps further testing) and the gathering of additional data were necessary.

The widespread use of testing in schools prompted leading test experts to caution against the limitations and possible misuses of testing. Mort and Gates (1932) reported, "In the first years of educational testing a great many abuses developed which retarded the progress of the movement to its full usefulness. In the last decade, however, workers have been able more and more to view the objective test with the perspective necessary to a realization of its shortcomings. As a result, standard tests are to-day quite generally looked upon as useful devices the limitations of which are so well know that they need not be feared" (p. iii). They also made note that testing would invariably influence the curriculum. In the long run even the best of teachers tended to emphasize the phases of the curriculum which were tested. However negative the influence on the curriculum was, the elimination of testing was not appropriate for Mort and Gates because they maintained that valuable information regarding placement or which children might benefit from remedial teaching could be gleaned from tests. "One of the most important uses of tests is the diagnosis of the individual needs of boys and girls on a broader basis than that which is generally understood by discovering the need of remedial teaching" (Mort and Gates, 1932, p. 8).

Hawkes, Lindquist, and Mann (1936) argued that "the first uses ordinarily assigned to achievement examinations, namely, the maintenance of standards and selection. . . are undoubtedly the dominant ones in most current examining" (p. 457). And they also added that another key use for tests was as an incentive to study. "Genuine educational motivation is a compound of the individual pupil's abilities and effective interests, and no motivation can long persist which is not fed by consciousness on the part of the student that what he is doing is significant. . . ." (Hawkes, Lindquis, and Mann, 1936, pp. 455-456.). They claimed: "Nevertheless it is perfectly clear that we are not going to abandon examinations. They are

necessary instruments of too many important educational and social purposes" (Hawkes, Lindquist, and Mann, 1936, p. 462).

Even though testing was deemed to be important, Cattell and Moodie (1936), stressed a support of using multiple criteria to make decisions about students as well. They maintained that if a performance test gave a lower mental age than the intelligence test, the examiner should have given another test. If there were discrepancies, then effort must be given to find out the true score so that the validity of the measure was not called into question. The test that gave the discrepant score was the test that was not appropriate (Cattell and Moodie, 1936).

The call for using multiple sources of information about a student continued. Tyler (1944) proposed that "the greatest advances can now be made in admission and placement by the use of examinations, although we recognize the importance of stimulating schools and colleges to maintain more adequate records which include observations, samples of work, and other evidence of student abilities, interests, and accomplishment" (p. 11). Tyler echoed the sentiment that more than one criterion should be used to make significant decisions about placement or admission for a student. A. E. Traxler further supported Tyler's view. Traxler (1944) held, "A single test score or other observation may be vague in meaning or, on occasion, even misleading, but its meaning becomes increasingly clear as the number of scores or observations of the same kind is multiplied. . . The cumulative record, therefore, should present not only a complete picture of a pupil's test scores over a period of years, but it should include a comprehensive summary of all the information that the school has about the pupil" (p. 30).

According to Lehmann (2004), Ralph Tyler believed "that commonly used standardized achievement tests did not provide a valid measure of what children have learned but were (and are) designed to rank students" (p. 26). Lehmann also identified three problems that Tyler saw

with standardized tests. The first problem that Tyler asserted was the purpose of standardized tests which "was to identify individual differences in achievement, not to measure individuals' learning" (2004, p. 26). Second, Tyler did not believe that the scores were reported in a manner that was valuable in determining the "achievements of a community" and, third, that the dependence on grade level norms "assumed some consistency within, as well as across, grade levels" (Lehmann, 2004, pp. 26-27).

According to Remmers and Gage (1943), the uses of educational measurement narrowly ranked students. Remmers and Gage (1943) professed the following:

Given the instruments, i.e., the achievement tests, with which to make such rankings the teacher or administrator may then use them for the following purposes:

1. To maintain standards
2. To select students
3. To motivate learning
4. To guide teaching
5. To furnish instruction
6. To appraise teachers, teaching methods, books, curricular content, etc. (p. 5).

Cronbach (1949) claimed, "Tests aid in making many sorts of decisions, including selection and classification of individuals, evaluation of educational or treatment procedures, and acceptance or rejection of scientific hypotheses" (p. 23). According to Bauernfeind (1978), Cronbach was sure to make the point that tests should be just one part in the decision making process for a student. Cronbach (1949) stated, "In sound practice, evaluation of a pupil or a teacher is never based on these tests alone; instead, the tests are treated as one source of data to be linked with many other facts in making a final evaluation" (p. 273). Bauernfiend (1978)

stated, “Cronbach further points out that tests can be used to aid four types of decision-making processes—selection decision, classification decisions, evaluation of treatments, and checking on scientific hypothesis” (p. 4). The key word in this passage was “aid.” Cronbach did not advocate the use of a single test score to make these four kinds of decisions about a student as the trend in public policy of the twenty-first century often dictates. “In sound practice,” Cronbach (1949) emphasized, “evaluation of a pupil or a teacher is never based on these tests alone; instead, the tests are treated as one source of data to be linked with many other facts in making a final evaluation” (p. 273). Cronbach (1949) also was sure to point out that when students want to do their best and understand the reasons for the testing, then accurate interpretation of results is possible. However, “[w]hen this is not the case, scores are invalid” (p. 268).

Cronbach (1949) interestingly pointed out as well that when the test becomes the goal of the entire year’s worth of work, then “the tests became a taskmaster which everyone in the school found himself trying to serve” (p. 273). Instead of assessments being an instrument created by and used by a teacher to determine how much his or her students had learned, the tests became the focus of everything that happened in the classroom, and “useful classroom activities which would not raise test scores received scant encouragement” (Cronbach, 1949, p. 273).

When the accountability movement of the 1960s emerged, Remmers, Gage, and Rummel (1965) cautioned, “avoid using tests to punish pupils or to foster a spirit of rivalry among pupils, teacher, or schools. Teachers and administrators must keep the welfare of pupils uppermost and be sensitive to the requirement of adequate human relations” (p. 175). In the late 1960s, Ralph Tyler worked steadfastly to help create an assessment that would help to determine how much students in the United States were achieving in order to better inform practice. It was Tyler’s

involvement if the development of the National Assessment for Educational Progress that began to help accomplish that goal. Epstein (1996) maintained, "When the late Ralph Tyler spoke in defense of a national assessment, he did not envision that one day the National Assessment for Educational Progress (NAEP) would become an instrument of accountability used to stimulate national reform of education" (p. 22). Early in the process, there were critics such as David Goslin who feared that the results of an assessment like the NAEP would have "'harmful effects on school curricula'" and would lead to "'misinterpretation of the results by the general public'" (Lehmann, 2004, p. 31).

Lehmann (2004) pointed out that Tyler was highly critical of the ways in which the results of standardized tests were being used in the 1960s. Tyler highlighted three main misuses of test scores: identifying individual differences in achievement rather than individual learning, the lack of meaningful test scoring and reporting in terms of the achievement within a community, and the faulty assumption that grade-level norms were consistent within and across grade levels (Lehmann, 2004). Tyler envisioned NAEP as "an assessment that would support teaching and learning, rather than select and sort students. Tyler was pretty scathing about a system focused primarily on grading and sorting (Mosher, 2004, p. 329).

The call for schools to be held accountable continued to be heard throughout the 1970s and 1980s. Regarding accountability, Tyler (1971) stated, "The limitations of standard achievement tests for this purpose are now being widely recognized. They do not measure what the pupil has learned but rather where he stands on a scale that arranges those who have taken the test from the highest score to the lowest" (p. 4). Additionally, according to Popham (2000), ". . . a meaningful amount of what's measured by today's high-stakes tests is directly attributable not to what students learn in school, but to what they *bring* to school in the form of their families'

socioeconomic status or the academic aptitudes they happened to inherit" (p. 18). How can those types of factors be controlled to get accurate and valid scores?

Cremin (1976) claimed that there need to be better methods for measuring educational accomplishments that what existed at the time. "For all our sophistication in testing—and we have made tremendous strides in the last decade or so—our instruments are still imprecise about what what should be evaluated and to what purpose. . . And they tell us next to nothing about where anything in particular has been learned. . . ." (pp. 88-89). But even more recently, L. B. Resnick (1981) asserted "today's standardized achievement test, even when accompanied by complex scoring services intended to reveal details of individual children's response patterns, do not respond to teachers' needs for information that can be used in planning instruction for small groups or individual children" (p. 624).

Reilly and Lewis (1983) had very specific uses of standardized tests in mind in their text entitled *Educational Psychology*. They made reference to the fact that the 1970s brought a strong stand against standardized testing from many based on the fact that those opponents felt that the tests were biased against certain groups of students. Reilly and Lewis (1983) identified the following as guidelines for the uses tests:

1. Standardized tests should be used only when there is specific reason for doing so.
2. Unless the scores from a standardized test are actually used for some specific purpose, the test cannot really be justified.
3. Whenever possible, students should be given feedback of the results of any standardized tests that are given. (p. 523).

These guiding principles allowed for the explicit use of scores to reach a certain end rather than just administering assessments for the sake of the administration. Additionally, timely and informative score reports assisted students and parents in determining the significance of the test results. Reilly and Lewis (1982) also had a specific list of ways that teachers could use test scores in planning: checking on content emphasis, individualizing instruction, grouping students, counseling students, identifying special needs, and measuring academic progress.

Even though the tests were seen as objective measures that enforce accountability, the results were "inaccurate, inconsistent, and biased against minorities, females, and students from low-income families" (Neill and Medina, 1989, p. 689). Neill and Medina (1989) held, "Instead of promoting accountability, tests shift control and authority into the hands of an unregulated testing industry" (p. 689). And since the judgments that educators make about children based on the scores are ones that have not been validated by the test authors and publishers, the judgments are "risky" (Neill and Medina, 1989, p. 691). According to Neill and Medina (1989), one of the most "devastating" uses of test scores was to determine whether or not a child was ready for first grade (p. 693). "Standardized tests for young children are among the least valid and least reliable exams. . . ." (Neill and Medina, 1989, p. 693).

By the 1980s, most scholars agreed that "the results of using one test administered one time as the primer determinant . . . have been so demonstrably fallacious" (Deighton, p. 184) that the practice should not be employed by responsible educators and policymakers. Using a single test score for making high-stakes decisions for students was not supported by most writers in the field of educational assessment and educational reform. In fact, the greatest criticism was that the scores were not used as one piece of information that will help to make decisions regarding a student. Clifford (1984) attributed some of that to overuse: "So widely are tests used that

misuse is common" (p. 389). Madaus (1988) declared, "[T]here are other uses of test results that do not always immediately and directly affect students but nonetheless are generally perceived by people as involving high stakes. For example, SAT results are of secondary importance in admission decision for those colleges trying to fill vacant seats in the face of adverse demographics. Nonetheless, individuals and school systems act on the perception that these college admissions test are of crucial and singular importance" (p. 87). Madaus (1988) further stated, "Testing programs should, in my view, be seen as an ancillary tool of curriculum and instruction,--albeit, a very necessary, useful, and important one—and nothing else. The long-term negative effects on curriculum, teaching, and learning of using measurement as the engine, or primary motivating power of the educational process, outweigh those positive benefits attributed to it" (pp. 84-85).

Madaus (1988) listed seven general principles that describe the impact of high-stakes tests:

> **Principle 1.** *The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon: if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false—the effect is produced by what individuals perceive to be the case.*
>
> **Principle 2:** *The more any quantitative social indicator is used for social decision making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.*
>
> **Principle 3:** *If important decisions are presumed to be related to test results, then teacher will teach to the test.*

Principle 4: *In every setting where a high-stakes test operates, a tradition of past exams develops, which eventually de facto defines the curriculum.*

Principle 5: *Teachers pay particular attention to the form of the questions on a high-stakes test (for example, short answer, essay, multiple-choice) and adjust their instruction accordingly.*

Principle 6: *When test results are the sole or even partial arbiter of future educational or life choices, society tends to treat test result as the major goal of schooling rather than as a useful but fallible indicator of achievement.*

Principle 7: *A high-stakes test transfers control over the curriculum to the agency which sets or controls the exam.* (italics in original, pp. 88-97).

These seven principles represented various considerations that test experts have given to the uses of test scores over the course of the twentieth century.

Madaus (1988) advocates that educators should work to make policymakers more knowledgeable about the "dangers associated with high-stakes testing" (p. 44). Madaus was also clear that the public has the right to hold systems, schools, and teachers accountable for student learning. However, he believed that we should "[negotiate] an agreement that a host of indicators of student achievement will be developed and used" (Madaus, 1988, p. 44). Unfortunately, the public continued to see tests as "an indicator of effectiveness" (Madaus, 1988, p. 35) simply because policymakers and the media used the results to rank schools.

The ranking of schools became the most important use of test scores in the eye of the public. The "[p]erceptions that a test has high-stakes associated with it are the ignition for test preparation and measurement-driven instruction" (Madaus, 1988, p. 36).

Using the content of a test to determine curriculum and therefore to drive instruction in schools was called Measurement Driven Instruction. This idea stemmed from the belief that what was tested on a particular assessment determined what teachers would teach and what students would learn. This was particularly true if the test was used to rank schools. The debate over whether or not Measurement Driven Instruction was beneficial to students reached a fever pitch in the late 1980s. Madaus (1988) asserted, "Proponents of testing argue that the power of testing to influence what is taught, how it is taught, what is learned, and how it is learned is a very beneficial attribute. This view of testing and curriculum is sometimes referred to measurement-driven instruction" (p. 84). Madaus (1988) stated that this type of instruction only focuses on the style of the test; that is, multiple choice questions about the most basic of skills. He maintained that testing and Measurement Driven Instruction "constrains the creativity and spontaneity of teachers and students and finally demeans the professional judgment of teachers" (p. 85). Madaus (1988) professed, "Measurement-driven instruction invariably leads to cramming; narrows the curriculum; concentrates attention on those skills most amenable to testing (and today this means skills amenable to the multiple-choice format)" (p. 85). Bracey (1987), a well-known critic of Measurement Driven Instruction, contended that this philosophy forces teachers to look at the knowledge and skills of the curriculum in isolation.

The supporters of Measurement Driven Instruction maintained that if more thought-provoking tests were given, then teachers would improve their instruction because it would match the more authentic, complex assessments (Firestone and Schorr, 2004). Popham (1987) called it "the most cost effective way of improving the quality of public education in the United States" (p. 679). Popham (1987) maintained that if MDI is "properly conceived and implemented" (p. 680), then it could be a positive force in the reform of public education. Part

of Popham's version of "properly conceived and implemented" included having curricular goals that were significant. He believed that the focus of instruction should be limited and that the assessments must be properly constructed so that they were not considered only after the instruction is over. Instead, the assessment should be created before the instruction was planned so that the instruction moved students toward the skills and concepts that were the desired learning targets.

Test experts asserted that the desired learning targets, or the curriculum, should be determined by educators and not left up to test publishers. Instead, proponents of testing determined that tests were beneficial in defining learning targets. They called it curriculum alignment. Shanker (1990) argued, "The very term 'curriculum alignment' is a fancy way of saying that tests narrow and determine the curriculum" (p. 5). While the debate on (MDI) continued, its opponents asserted teachers engage in "decontextualized test preparation. . . [which] is a special activity only loosely related to [the teacher's] regular lessons and focused on the test itself" (Firestone and Schorr, 2004, p. 2). Kreitzer and Madaus (1985) declared, "Where proponents of high-stakes testing see focused curricula, opponents see dangerously narrowed curricula. Where proponents see rising scores, opponents see misleading scores devoid of meaning" (p. 26). The idea that assessments will broaden the emphasis of the curriculum by asking the right kinds of questions seemed illogical to Kreitzer and Madaus. Noble and Smith (1994) agreed when they held, "policy makers and scholars who still believe in the power of assessment to drive reform and change schools have focused on the fallacies in the psychology and pedagogy of the traditional view as well as the form of the measurement itself" (p. 3). It is not the test itself that will bring reform to the curriculum and to instruction in the United States.

Madaus and Kellaghan (1992) were certain in writing, "Much of the testing that goes on today (particularly mandated, high-stakes testing), its sponsorship, financial base, character, and use, is also essentially bureaucratic and only secondarily educational, or if it is educational, it is educational as conceived by policymakers" (p. 121). Ansley (2000) is very clear about what has happened in the arena of school accountability and standardized testing over the last few decades:

> These tests, like many other aspects of education, have become pawns in a political chess game. In most such states, these tests are transformed from evaluation devices to high stake accountability tools. . . . This is a large departure from the purposes for which these tests are constructed. (p. 278).

Epstein (1996) claimed, "The results of high stakes standardized tests provide the illusion of begin objective measures of the effectiveness of schools. However, most of these tests have been limited to what has been easy to measure, that is, basic skills, rote factual information, and lower levels of cognitive thinking" (p. 32). Epstein (1996) stressed that those who promote such high-stakes testing see it as "the one mechanism capable of raising the standard of education" (p. 32). But what many tended to overlook was that when tests are not used for appropriate purposes, they lose validity.

Test validity in terms of the purposes for which the tests were used was not the only concern raised. The varying demographics of the schools that were assessed was a problem as well. "[A]t least one study found that 89 percent of the variation in state average test scores in the NAEP 1992 Trial State Assessment in mathematics can be explained by the combined effects of four demographic variables.' Because of the uncontrollable nature of such variables, the study

concluded that NAEP results should not be used 'for the purpose of *comparing* and *ranking* states according to the relative quality or proficiency of the states' educational programs' (Robinson and Brandon, 1994, pp. 15-17)" (Epstein, 1996, p. 33). One of the long-time fears that educators had regarding NAEP was that it would be used for accountability and not just a progress monitoring assessment. L. B. Resnick (1999) echoed this concern:

> Using NAEP—or a test closely based on it—to track the performance of individual schools and districts would convert NAEP from a monitoring to an accountability instrument. This would create a national presence in American schools far greater than anything we have seen before. The NAEP tests would be much more influential and constraining, for example, than requirements for Title I, special education, or Goals 2000 expenditure. District- and school-level score reporting would give NAEP influence over the de facto curriculum: that is, over what is taught day-to-day and especially what is taught close to test-taking time. (p. 3).

L. B. Resnick (1999) emphasized that we cannot put an assessment in the classroom that acts as a "thermometer" and not expect changes in that classroom. "But teachers and school principals who are held accountable will produce efforts on their part to have their students perform well on that assessment. Primary among these efforts will be teaching the test item—that is, having students practice doing the very things that will appear on the test" (L. B. Resnick, 1999, p. 6).

According to Jones' (1996) history of the NAEP, examining student achievement was the original purpose of the assessment. He also asserted that the assessment should continue in it tradition of exploring trends and that since this practice was "not compatible with a high-stakes

accountability assessment program, NAEP should avoid serving this added purpose (Jones, 1996, p. 15).

The media and policymakers have insisted that test scores are acceptable ways to hold schools accountable. Parents and the public "view these [test] scores in isolation, leading to a troublesome overemphasis of the usefulness of these scores" (Ansley, 2000, p. 270). Michael Apple (2000) stressed the following:

> This concern for external supervision and regulation is not only connected with a strong mistrust of producers (e.g., teachers) and to the need for ensuring that people continually make enterprises out of themselves; it is also clearly linked both to the neoconservative sense of a need to return to a lost past of high standards, discipline, awe, and real knowledge and to the professional middles class's own ability to carve out a sphere of authority within the state for its own commitment to management techniques and efficiency. (pp. 65-66).

And Ansley (2000) also asserted that local boards of education should be interested in their schools' scores on a battery of tests as "a single piece of a fairly large and complex puzzle of educational achievement" (p. 279). Scholars maintained that teachers taught "to the test" in an effort to improve student performance on exams and that this unduly influenced the curriculum in ways that were not entirely positive. Ansley (2000) was also critical of using standardized test scores to determine the effectiveness of individual teachers. "This clearly represents a gross misuse of test scores" (p. 279). Scores that might indicate that students are low-achieving did not necessarily mean that the instruction that those students receive was poor. Additionally, schools that had really high test scores were by no means above improving their

instruction in some manner (Ansley, 2000). "Today, this practice of singling out low-scoring schools to urge their instructional staffs to shape up 'unacceptable' performances is incredibly widespread" (Popham, 2001, p. 17). Interestingly, Popham's view that high-stakes tests are not appropriate instigators of school reform was a change from his views published in the 1980s and cited earlier in the study.

Popham (2004) claimed that this "shaping up" of teachers and schools often takes the form of narrowing not only the curriculum but also how the teacher instructs her class. Popham (2004) contended, "First, because of substantial pressures to raise students' scores on high-stakes tests, in many instances we find educators abandoning significant curricular content not measured by their local high-stakes tests. . . . Content not assessed on a high-stakes test is content cast aside" (p. 65). This has been one of the most valid criticisms of the testing movement. Educators deemed such standardized tests as essentially robbing the classroom teacher of the ability to determine what has been the knowledge of most worth in her classroom (Deighton, 1971, p. 183).

The American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2007) jointly published *Standards for Educational and Psychological Testing* that called for test users to be sure that they were utilizing the scores of an assessment in an appropriate manner. Over and over, test users are cautioned that the validity of a particular test no longer exists when the results are misused and/or abused for purposes other than what the test authors and publishers intended. Additionally, the Joint Committee published guidelines for informing test users, for test users' information in selecting tests, and for informing test takers. However, the Guidelines from the Code of Fair Testing Practices in Education regarding developing and selecting appropriate tests

were relevant to this study and are as follows (The Joint Committee on Testing Practices, The American Psychological Association, 2007):

Test developers should provide the information and supporting evidence that test users need to select appropriate tests.

1. Provide evidence of what the test measures, the recommended uses, the intended test takers, and the strengths and limitations of the test, including the level of precision of the test scores.
2. Describe how the content and skills to be tested were selected and how the tests were developed.
3. Communicate information about a test's characteristics at a level of detail appropriate to the intended test users.
4. Provide guidance on the levels of skills, knowledge, and training necessary for appropriate review, selection, and administration of tests.
5. Provide evidence that the technical quality, including reliability and validity, of the test meets its intended purposes.
6. Provide to qualified test users representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports.
7. Avoid potentially offensive content or language when developing test questions and related materials.
8. Make appropriately modified forms of tests or administration procedures available for test takers with disabilities who need special accommodations.
9. Obtain and provide evidence on the performance of test takers of

diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed.

One of the most important of these guidelines was the first one that called for the publisher or author of a test to give the recommended uses for the assessment.

Regarding the interpretation of test results the Joint Committee (2007) recommended that the following guidelines be followed:

Test developers should report test results accurately and provide information to help test users interpret test results correctly.

1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.
2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.
3. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information

about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.

4. Encourage test users to base decisions about test takers on multiple sources of appropriate information, not on a single test score.
5. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.
6. Specify appropriate uses of test results and warn test users of potential misuses.
7. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.
8. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.

These guidelines provided the testing community with the actions to which responsible test creators should adhere when developing and marketing tests.

Summary

This chapter has reviewed the works of scholars in the field of educational measurement and what they have deemed as the appropriate or inappropriate uses of test scores. The appropriate uses include the following:

1. As one source of information regarding student strengths and weaknesses
2. As one source of information for placement or admission into certain educational programs

3. As one source of information to judge teacher or program effectiveness. The inappropriate uses identified by the leading test designers are as follows:

1. As a sole source of information for promotion, retention, graduation, or other high-stakes decisions about students

2. As a sole source to rank schools

3. To foster competition among schools

4. As a sole source of curriculum

5. As a substitute for teacher judgment regarding student achievement

From the early works of Alfred Binet to the more recent work of the American Psychological Association, there were repeated cautions against using a single test score to make high-stakes decisions about students such as promotion, retention, and graduation. Notables in the field such as L. J. Cronbach cited that tests should be just a part of the decision-making process. The practice of using test scores to rank schools and to foster competition was repeatedly condemned by the likes of Ralph Tyler. Additionally, detractors of such uses of standardized tests maintained that the test then became the curriculum and that the tests were never intended to be used for such purposes. Examinations that determine curriculum were believed to no longer be valid measures of student achievement. It was interesting to note that from the beginnings of the testing movement until present day, scholars have agreed that test scores alone should not be used to judge teachers or their ability, just as their opinions of student achievement should not be dismissed in lieu of what a multiple choice test said about a student.

Chapter 5

THE APPROPRIATE AND INAPPROPRIATE USES OF STANDARDIZED TEST SCORES: TEST PUBLISHERS

This chapter outlines the test publishers' identified appropriate and inappropriate uses of test scores. Assessments reviewed are the Stanford Achievement Test, the Scholastic Aptitude Test, the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the California Achievement Test, the Comprehensive Test of Basic Skills, and the National Assessment for Educational Progress. Interpretive guides and manuals for each of the examinations were analyzed when they were available. Validity of test scores when they are used for specific purposes as defined by the documents issued by the test publishers.

The Stanford Achievement Tests

According to the 1926 Stanford Achievement Test Manual of Directions, the intention of the authors of the assessment was to remedy for Grades 2 through 8 tests "that [are] fragmentary in the sense that [they] cover only a part, and often only a small part, of the ground that the average teacher or administrator desires to cover" (Ruch and Terman, 1926, p. 3). The manual also cited that the tests of the day are based on norms that were "derived of such diverse methods that there is no satisfactory way to compare a pupil's score in one subject with his score in any other, or to summate a pupil's scores for the various subjects into a composite score" (Ruch and Terman, 1926, p. 3). The manual went on to assert that the score received by a student on any subject test can be appropriately compared to his score on another subject test within the same administration (Ruch and Terman, 1926).

Ruch and Terman (1926) identified as one of the primary functions of the Stanford Achievement Test the ability to use the assessment to sort students who are entering high school, as "[e]xperience has shown that better results are obtained when these diverse abilities are grouped into relatively homogeneous sections or classes" (p. 4). The authors went on in the manual to point out that the intention was to create a battery of tests that addresses all of the curriculum for Grades 2 through 8 and that it be an examination that would be easy to administer in a reasonable amount of time with consistent and reliable results (Ruch and Terman, 1926). Ruch and Terman (1926) maintained the following:

> We are no longer content with tests so rough that they are useful only for comparing one school or one city with another. We now demand that a test shall give a dependable measure of the individual pupil, in order that we may use his score for placing him in the grade where he belongs. This is the most important function of standard tests of every kind, a function which requires that the probable error of a score shall be a relatively small fraction of the increment between successive grade means. (p. 7).

The test as a way to appropriately place individual students was the goal rather than comparing students to one another within schools or from school to school. "It is assumed that the Stanford Achievement Test will be given in order that the results may be used, and not merely to gratify and idle curiosity as to how the school stands with reference to other schools" (Ruch and Terman, 1926, p. 54).

However, one function of high stakes tests that is often seen today, determination of promotion or retention, is a purpose for which Ruch and Terman (1926) thought the 1925

Stanford Achievement Test was appropriate. But the authors cautioned users of the assessment results that if the daily work of the student was not comparable to how they perform on the assessment, a closer look must be taken to determine the validity of the test results. Instead of the fault lying in the assessment, the authors felt that, more often than not, the problem was that the classroom teacher was giving high grades to a student whose personality was appealing over that of a more shy, withdrawn student.

Ruch and Terman (1926) also recommended using an intelligence test such as the Binet to discover why there might be a discrepancy in a student's score and his everyday work. Regardless of the reason, the authors advised that the single score not be the sole basis for decisions about the student and that when there were significant gaps, the cause must be determined for the good of the student (Ruch and Terman, 1926). "It may be advisable to use a group test of intelligence along with the *Stanford Achievement Test* for purposes of classification" (Ruch and Terman, 1926, p. 4). The authors further stated, "It is important to note any marked discrepancy between a pupil's test score and the apparent quality of his daily work" (Ruch and Terman, 1926, p. 57). The authors even cited such factors as "late entrance, irregular attendance, lack of interest, poor application, or poor teaching" (Ruch and Terman, 1926, p. 58) for scores that were seemingly too low. They clearly intended that educators should work diligently to help students reach their potential when there was a considerable difference between what the student ought to be able to accomplish based on intellect and what the student is actually accomplishing based on daily work and their performance on the Stanford Achievement Tests.

In order to help test users employ test scores appropriately, the publisher began to provide documents designed to aid them in the interpretation of test results. With each new

edition of the Stanford Achievement Tests, there was published a Technical Data Report. The Technical Data Reports from 1985, 1990, 1997, and 2004 were reviewed. The 1985 Technical Data Report described the development of the seventh edition of the Stanford Achievement Tests. The 1989 report corresponded to the eighth edition of the assessment. In 1997, Harcourt Brace Educational Measurement published the Technical Data Report for their Ninth Edition of the Stanford Achievement Test. Finally, the 2004 version of the report informed users about the tenth and most current edition of the assessment. The manuals were very similar to each other with only minor changes made from one edition to the next. They reported that there are various kinds of scores that are available from the Stanford Achievement Test. Those scores include scaled scores, individual percentile ranks, stanines, normal curve equivalents, grade equivalents, achievement/ability comparisons, group percentile ranks and stanines, and content cluster and process cluster performance categories.

The Technical Data Reports asserted that the "the particular scores to be used depend on the purposes for which the test has been given" (1985, p. 22; 1990, p. 37; 1997, p. 32; 2004, p. 34). The reports stated that "scaled scores [are] especially suitable for comparing results when different forms for levels of the test have been administered and for studying change in performance over time" (1985, p. 22; 1990, p. 37; 1997, p. 32). The 2004 edition of the report claimed, "Scaled scores are especially suitable for comparing student performance in a particular subject area over time" (p. 34). The individual percentile ranks were assigned value "when position in the reference group is of primary interest" (1985, p. 24; 1990, p 41; 1997, p. 34; 2004, p. 35). The uses of the other available scores are identified to be useful for comparing aspects of subtests within the battery of the Stanford Achievement Tests (Technical Data Report, 1985, 1990, 1997, 2004). While the reports were sure to indicate that the various scores reported had

specific uses, making high-stakes decisions such as promotion or retention were not included in the analysis of uses of test scores.

Stoker (1992) compared the *Code of Fair Testing Practices* to the guide provided by the publisher. He cited the part of the *Code* that stated "Test developers should: Warn users to avoid specific, reasonably anticipated misuses of test scores. Warnings appear in more than one Stanford booklet" (Stoker, 1992, p. 865). Stoker (1992) was explicit that the *Guide for Organizational Planning* that accompanied the SAT described the limitations of the test and that issues related to both the uses and the misuses of the test are discussed.

While much of the 2004 Technical Data Report of the Stanford Achievement Test, Tenth Edition, was very similar to the 1997, there were additional details. For example, the 2004 edition was very clear that the scaled scores "enabl[e] the comparison of students' test scores with those of other students and the evaluation of changes in student performance across subtests and testing occasions" (p. 34). The report maintained that "content clusters and subclusters can be useful in identifying students' strengths and weaknesses within a broader content area" (2004, p. 42).

In addition to the 1985, 1990, 1997, and 2004 Technical Data Reports, the Guide for Classroom Planning for seven different levels of the 2003 Stanford Achievement Test were also published. The levels were as follows: Advanced ½, Task 1/2/3, SESAT ½, Primary 1, Primary 2, Primary 3, and Intermediate 1/2/3. In each guide, there were constants that were stressed in terms of using the test scores. Each guide contained a section entitled "How Did My Students Do?" This section in each guide opened with the same paragraph (Intermediate 1/2/3, 2003):

> The results of a norm-referenced achievement test can be of great value to teachers when considered together with information from other sources.

> Test results can give you information about specific areas of strength and weakness in achievement for individuals or groups of students; help you set up instructional priorities; assist you in grouping students for instruction; and allow you to compare the performance of your students to national norms. Remember, however, that test results are a picture of a student's achievement at a single point in time. Test results must be considered in light of performance in bother large- and small-group activities, informal assessments, teacher observations, and checklists, portfolios, and logs. (p. 64).

The general guidelines in all of the guides called for teachers to do the following when looking at test scores: "Watch for the unusual," "Always ask 'why?'," "Don't overemphasize small differences," "Interrelate information from different subtests," and "Don't expect to discover something new and different about every student" (Guide for Classroom Planning Primary 1, 2003 pp. 46-47). Additionally, the guides revealed that test users should keep many factors in mind when using test scores. "As you examine and interpret students' test results, remember that achievement in school and on the test may be affected by any of the factors listed below. . . ." (Guide for Classroom Planning Advanced 1/2, 2003, p. 61). Those factors included student health, home environment, student age, school attendance, interest, and study and work habits. School factors such as expectations and level of instruction were also cited as factors worthy of consideration (Harcourt, 2003).

The Guides for Classroom Planning (2003) each had a section entitled "Where Do I Go From Here?" that was nearly identical for each test level as well. This section in each guide called for the teacher to set instructional priorities and gave the steps for doing so with the test results in mind. Then, this section emphasized how test results might be used for grouping for

instruction and monitoring progress of students. Finally, these sections insisted on sharing results with parents in order to form a partnership (Harcourt, 2003). Harcourt Assessments did not intend for the single score on any level of the Stanford Achievement Test to be used to make important decisions about an individual student. The goal of the publisher was for the test results to be one of several factors in determining the next steps for students and teachers. The guides did not recommend the comparisons of schools and school districts and determinations of schools' effectiveness.

The Scholastic Aptitude Test

The purpose of the Scholastic Aptitude Test had been clear from the beginning: "to supplement the school record and other information about the student in assessing his competence for college work" (Donlon and Angoff, 1984, p. 15). Donlon and Angoff (1984) also asserted that the test can give information about students that cannot be gleaned from any other source. In fact, they stated that the assessment was originally intended to highlight inconsistencies in a student's secondary school record so that they could be addressed in an effort to eliminate inflated grades because a student was a teacher pleaser (Donlon and Angoff, 1984). In 1925 that a committee headed by Carl C. Brigham created a manual for the Scholastic Aptitude Test (SAT), which was meant to be a test of whether or not students can generalize facts to apply to a variety of situations and not just regurgitate them. The public and those who would use the test scores for evaluating a student's knowledge and skills were warned by the committee that the SAT should be merely a "supplemental record" (Donlon, 1984, p. 2). "The manual further declared that 'to place too great emphasis on test scores is as dangerous as the failure properly to evaluate any score or rank in conjunction with other measures and estimates which it supplements'" (Valentine, 1987, p. 35).

Calvin (2000) maintains that Chauncey had no desire to create "an elite based on financial wealth and birthright" (p. 5). The 1926 manual that the College Board issued contained a warning that the SAT "should be regarded merely as a supplementary record (Angier et al., p. 1) was the basis for the many documents that followed it. In a memo from Frank Bowles (1949), Director of the College Board, written to principals and headmasters, Bowles claimed that many of the members of the College Board were "apprehensive. . . that undue emphasis will be put upon scores (scores are only one of the criteria used in guidance and admission." The two volumes of *College Board Scores: Their Use and Interpretation* in 1948 and 1949 and the yearly *Description of the SAT* published yearly from 1956 through 1971 communicated the same the message: the scores of the SAT are intended by the College Board to be used as one component in determining a student's suitability for college. "It is, therefore, neither feasible nor desirable to outline a single ideal method of using the College Board test for admission or to suggest that all the colleges ought to use it. . . Any system for selecting students wisely, however, depends ultimately on two types of information—a reasonably accurate description of the applicant as he is, and a reasonably accurate prediction of the kind of college student he is likely to become" (Dyer and King, 1954, pp. 19-20). Each of these documents, along with in its 1962, 1963, 1964, 1965, 1966, and 1967 *SAT: A Guide for Counselors and Admissions Officers*, encouraged colleges and universities to use the secondary school record along with the student's score from the SAT in order to determine whether or not a student was admitted to a particular institution or to make placement decisions regarding that student once the student is admitted. Similarly, each document stated that secondary schools should use the preliminary score on the SAT (that score earned before the senior year in high school) as a guidance tool.

The College Board in its 1957 through 1960 annual *Candidates and Tests* identified the students who take the SAT as "'preliminary'" candidates, those taking the test for guidance purposes, and as "'final'" candidates, those taking the test for college or university admission (p. 1). And in its 1962, 1963, 1964, 1965, 1966, and 1967 *SAT: A Guide for Counselors and Admissions Officers*, the College Board asserted that the SAT should correlate closely to the grades that the student received in secondary school. No publication from the College Board or from Educational Testing Service advocated that the SAT scores be used to compare students, schools, school systems, or states. While the increase in the numbers and kinds of tests that the College Board was giving made it seem that everything was going very well for its flagship test, the SAT, that is not the case. In 1963, the mean SAT scores began a decline that alarmed educators. Even though the reasons for the decline were complex, secondary schools began to get all the blame (Donlon, 1984). The College Board came to the defense of the secondary school system in the United States and repeatedly cautioned that this was not the proper use of SAT scores. They reiterated that the success of schools and the evaluation of students could not be based solely on one test score (Donlon, 1984).

The College Entrance Examination Board again in the 1970s "issue[d] repeated warnings that the SAT is not intended as a measure of a school's educational accomplishments" (Donlon, 1984, p. 5). Additionally, the College Board specifically opposed the use of the scores to compare states (Cameron, 1989), but it suggested using the scores to compare students within a state. "Although the College Board cautions against state-by-state comparisons because the percentage of high school graduates taking the SAT varies widely by state (from 3 percent to 69 percent), trends over time within a state can reveal how its students are progressing toward educational goals" (Cameron, 1989, p. 7). This use of test scores to compare students within a

state to determine if they were making improvements over a period of time was advocated by the College Board while they warned against comparing states to each other.

The Iowa Test of Basic Skills

Peterson (1983) asserted the following:

> On the practical uses of test results: "To the school administrator or supervisor, therefore, the Academic contest provides a unique source of important information about his own school. It furnishes him with a reliable means for evaluating the quality of instruction in his own school, as well as for checking the validity of the content taught; it helps him to discover those teachers on his staff most in need of supervisory aid, those subjects most in need of curriculum revision, and those pupils most in need of individual attention; it increases the reliability of the marks used for the promotion and demotion of pupils; it makes possible better educational guidance, and it gives him an indirect measure of the effectiveness of his own organization and administrative policies." (p. 13).

These earliest Interpretive Guides for the Iowa Test of Basic Skills were called Subject Matter Circulars. They advised that school administrators use the scores to evaluate the curriculum and the personnel within their own schools. They also recommended the use of the score to identify areas of weakness for students so that proper instruction could be given to students based on need. The guidance that could be provided to a student based on the results of the assessment was considered to be very valuable indeed.

The University of Iowa College of Education's Iowa Testing Programs is quite clear on its website and through the Interpretive Guides for the Iowa Test of Basic Skills (ITBS)

published by Riverside. As the authors of the assessments, the Iowa Testing Programs (2007) stated that the results from the ITBS should be used for instructional planning. "When used as intended, such batteries can be a useful supplement to teacher observations about what students are able to do, and they can provide a starting point for monitoring year-to-year student development."

Despite efforts to contact the Iowa Testing Programs for additional Interpretive Guides, this researcher did not receive the materials requested. Therefore, a review of secondary sources was necessary in order to determine the content of those guides dated prior to 1993. Harris (1978) published a review of the assessment in *The Eighth Mental Measurements Yearbook*. In that review, Harris (1978) stated that the publishers claimed that the "battery can be used to diagnose specific strengths and weaknesses of individual pupils" (p. 55). Harris (1978) also quoted the manuals as saying that the tests could be used "'to determine the relative effectiveness of alternate methods of instruction and the conditions which determine the effectiveness of the various procedures'" (p. 55).

A review of the University of Iowa College of Education's website along with the *Interpretive Guide for Administrators, the Interpretive Guide for Teachers and Counselors*, and the *Guide to Research and Development* revealed that the appropriate purposes for testing as identified by the Iowa Testing Programs (author of the ITBS) had changed little. In a few instances, the order of the identified appropriate uses may have been changed, but the content remains relatively the same. The 1994, 1996, and 2003 *Interpretive Guide for School Administrators* for Forms A, B, K, L, and M are roughly the same. However for 2003, there were new inappropriate uses added to the *Interpretive Guide for School Administrators*. The additions include the caution not to use "only the scores from a single test or battery to identify

the ‘best’ schools in a state or region” (p. 13). An additional identified inappropriate use is “using the scores form a single achievement battery or test to evaluate the effectiveness of the instruction of a certain teacher” (p. 12). One can only deduce that such revisions to the guides are a result of practices that are becoming more prevalent as a result of the accountability movement among politicians and the public. The 2003 guide includes language about standards for the first time.

In the *Interpretive Guide for Teachers and Counselors* (2003), the section entitled “Appropriate Purposes” has changed somewhat. The format is different, and the information is presented more as prose rather than a list of suggested uses of the results. The language of the guide is more inclusive of other means that teachers need to use in order to determine the next steps for instruction for particular students. Specifically, the guide says that “some of the information that teachers need” should come from “achievement batteries” (p. 6). The guide confirmed that the ITBS should be used as one of many pieces of information that parents and teachers use to determine the progress of individual students. “When used as intended, such batteries can be a supplement to teacher observations about what students are able to do , and they can provide a starting point for monitoring year-to-year student development” (Hoover et al, 2003, p. 5). Additionally, each inappropriate use identified begins with “using scores from a single test or battery to. . .” or “using only the scores from a single achievement battery or test to. . . .” (Hoover et al, 2003, pp. 7-8).

The appropriate uses of the scores as identified by the Iowa Testing Programs in the *Interpretive Guides for Administrators, Teachers, and School Counselors* include the following guidelines (2003):

1. To help determine the extent to which individual students have the background and skills needed to deal successfully with the academic aspects of an instructional program or a planned instructional sequence;
2. To estimate the general developmental level of students so that materials and instructional procedures may be adapted to meet individual needs;
3. To identify the areas of greatest and least development to use in planning individual instruction for early intervention;
4. To establish a baseline of achievement information so that the monitoring of year-to-year developmental changes may begin;
5. To provide information for making administrative programming decision that will accommodate developmental differences;
6. To identify areas of relative strength and weakness in the performances of groups (e.g., clases), which may have implications for curriculum change—wither in content or emphasis—as well as for change in instructional procedures;
7. To provide a basis for reports to parents that will enable home and school to work together in the students' best interests. (p. 7).

The inappropriate uses of the scores according to the Iowa Testing Programs in the *Interpretive Guides for Administrators, Teachers, and School Counselors* are as follows (Hooever et al, 2003, pp. 8-9):

1. *To screen children for their readiness for school enrollment*. The skills measured by these batteries are sensitive to short-term individualized instruction. Consequently, deficiencies in any of them are more likely to be due to limited opportunity to learn or to slow verbal development than

to delayed emotional or social development. The results from an achievement battery should never be used alone to make such important placement decisions.

2. *To retain students at a grade level*. There is considerable disagreement among educators about the appropriateness of grade retention. If a retention decision is to be made, assessment data gathered by the teacher over a period of months is likely to be the most relevant and accurate basis for making such a decision. It should go without saying that test scores from an achievement battery should not be used alone, or even be given major weight, in making a retention decision.
3. *To evaluate the effectiveness of an early childhood program*. The amount of emphasis given to academic objectives in an early childhood curriculum varies substantially among schools. All programs give attention to students' cognitive, physical, social, and emotional development, but the balance among the curriculum components in any given school ordinarily will depend on the nature of the students' background experiences, the philosophy of the teachers and administrators, and the demands of the community. Since achievement batteries can assess only a limited part of the total curriculum, test scores alone cannot describe the relative success or effectiveness of the entire program. Especially for programs that maintain a nonacademic or play-centered curriculum for the early years, scores on achievement tests provide only partial information about program effectiveness.
4. *To decide which instructional objectives should be taught at a certain grade*

> *level*. The questions on each test of the battery are only a small sample from a very large number of questions that potentially could be asked. For example, the 29 questions on the Level 5 Vocabulary test represent a small fraction of the hundreds of words that could be presented to test the development of students' listening vocabularies. There is nothing so important about each of those 29 words that teachers ought to teach them to their students. In fact, such teaching would destroy our ability to use the test score to generalize about the extent of each student's vocabulary development. In sum, no test question deals with an essential element of knowledge; each question is only representative of a larger collection of important elements. (pp. 8-9).

The recommendations were consistent with appropriate and inappropriate uses of test scores that were espoused by scholars in the field and as described in Chapter 4.

The Metropolitan Achievement Tests

According to Pullias (1941), the manual for the Metropolitan Achievement Tests stated that "recommended administrative and supervisory uses include, 'To rate teacher effectiveness,' 'Comparison of achievement of school with school,' to obtain 'accurate estimate of the relative efficiency of schools. . . and of the administrators'" (p. 28). The authors of the assessment at the time clearly intended that the test results be used for accountability measures. The 1947 revision of the manual included details for how to formulate a testing program that gave "emphasis on the more constructive look that goes with fall testing" as a tool for instructional planning (Findley, 1953, p. 48).

When the fifth edition of the MAT was published, the intended uses for the results of the assessment included that "'[i]t is an instructional planning tool that provides detailed prescriptive

information on the educational performance of individual pupils in terms of specific instructional objectives'" (Linn, 1985, p. 967). Linn (1985) argued that that Teacher's Manuals provided "teachers with many suggestions of activities to help correct identified weaknesses" (p. 967). He also cited, though, that there was not enough support for the validity of such claims (Linn, 1985).

The 1988 version of the *Metropolitan Achievement Test Survey Battery Technical Manual* did not offer any specifics as to what the uses of the test scores should be. The Psychological Corporation was still listed as was the primary publisher of the assessment even though that company was a subsidiary of Harcourt Brace Jovanovich by that time. However, by the time the seventh edition of the Metropolitan Achievement Test (MAT 7) was published in 1993, Harcourt had added the same section to the Technical Manual that had been included in the technical manuals of the Stanford Achievement Test.

That section dealt with the types of MAT7 scores that could be obtained. "Since the underlying properties of these scores are not necessarily the same, the particular scores to be used depend on the purposes for which the test had been given" (Harcourt, 1993, p. 33). The types of scores that were available included scaled scores, percentile ranks, stanines, normal curve equivalents, grade equivalents, achievement/ability comparisons, functional reading levels, content cluster performance categories, proficiency statements, and predicted SAT and ACT performance ranges. The manual indicated that the scaled scores are "especially suitable for comparing results when different forms or levels of the battery have been administered and for studying change in performance over time" (Harcourt, 1993, p. 33). The percentile ranks were deemed "useful in obtaining an indication of the relative standing of a student in comparison with other students in the same grade tested at the same time of year" (Harcourt, 1993, p. 35).

The other scores were noted not to be useful in comparing student scores across tests but only within subtests of the same administration.

Regarding the MAT7, Finley (1995) stated that a section in the teacher's manual argued that teacher should not question small differences in scores and that they should "interpret in light of other factors" (p. 605). Finley (1995) also declared that the manual addressed "how to use test results mentioning the misuse of test results in promotion and retention of students, grading students, teacher evaluation, and comparison of different tests, as well as the proper use of test results in establishing instructional priorities and grouping for instruction" (p. 605). Hambleton (1995) specified, "The publisher was clear about four inappropriate uses of the MAT. These inappropriate uses include promotion and retention of students, grading of students, teacher evaluations, and comparison of result from different tests" (p. 606).

The California Achievement Tests

The California Achievement Test was developed "to provide test users with information to guide instruction and improve learning" (Nitko, 2006, p. 8). According to the California Test Bureau Manual for the California Achievement Tests Complete Battery (1951), the results of the assessment should be carefully examined if the student's scores were below a "desirable standard" (p. 8). In such cases, according to the manual, test item analysis provides the most pertinent information to informing next steps for a student. The main goal of the manual is to deliver the message that teachers should use the data to look at individual or even class progress. However, the manual cautions against using test items as the basis of instruction. The manual also cites the following as appropriate uses for supervisors, principals, and superintendents (California Test Bureau, 1951):

1. To refute ill-founded charges that school achievement is below reasonable expectations, when test results show achievement to be satisfactory

2. To determine whether differences in achievement between succeeding grades is satisfactory

3. To determine whether the objectives of the curriculum are being Achieved

4. To determine whether marking practices in various schools reflect the true performance of the students as revealed by the test results

5. To determine whether the proportion of student "failures" (where students are failed) reflects the true performance of the student as revealed by the test results

6. To use as a basis for developing policies on ability grouping of students for instructional purposes

7. To determine whether the achievement test results are reasonable and satisfactory in light of the intelligences of the student and other related factors. (p. 11).

A 1953 review of the CAT, Schindler (1953) asserted that the publishers' intent for the use of the scores was that of a diagnostic nature. In fact, Schindler called them "extensive claims which the author makes for the tests as diagnostic instruments" and further cited that those claims "are not well founded" (p. 6). In the same publication of the *Mental Measurements Yearbook*, Shores (1953), applauded the manual's care "against overstatements concerning desirable uses to which the result can or should be put" (p. 8). However, Shores (1953) pointed

out that the authors did indeed cite that diagnostic uses were appropriate for the results of the CAT.

The 1957 *Technical Report on the California Achievement Tests* stated that "the most meaningful single score derived from the measurement of achievement" (p. 35) was the Grade Placement. The manual dictated that the score should be used to determine a student's placement based on the grade level for which the score attained by a student is the average. However, the manual also stated that the percentile norms can be used when the Grade Placement subtest data are not consistent. The manual also discussed interpreting Anticipated Achievement scores. Anticipated Achievement "for a pupil is the norm performance of a nation-wide sample of pupils in the same grade with similar mental and chronological ages" (California Test Bureau, 1957, p. 36). However, there was no mention of how to use the scores to compare pupils, schools, or states. There was also no mention of whether or not the authors believed that a single test score could give an accurate representation of a student's achievement (California Test Bureau, 1957).

In 1961, the Research and Development Staff of the California Test Bureau prepared a book called *Questions and Answers*. In this book, they answered forty-three questions about the 1957 California Achievement Test. Among those questions, only a few dealt with test scores, and only one dealt indirectly with the use of test scores. The book called for mental ability, chronological age, and school-grade classification to be considered as important factors when interpreting the scores that students received on the California Achievement Tests. The only mention of promotion policy is on page seven when the authors related low test scores to promotion or retention in a given grade. Their logic was that students who are retained must not be of typical mental age because "pupils of normal mental ability should ordinarily progress

through the grades at the rate of one grade per year" (California Test Bureau, 1961). If a student who had been retained did not score well on the test, then it must have been because that student was of sub-normal mental ability.

The 1963 manual for the CAT described the use of the tests in this manner: "the norms 'may be considered as the test performance which the student would be expected to attain' [which] is more appropriate than the statement from earlier manuals that 'they may be considered as the test performance which the student should attain'" (Merwin, 1965, p. 19). The difference in the terminology from "should" to "would be expected" removed some of the judgmental tone from the manual. It also allowed for the individual differences in student achievement that were to be expected dependent upon a variety of factors such as mental ability and anxiety level. In any case, the 1963 or earlier manuals did not identify the use of a single CAT score to make decisions such as promotion, retention, or placement as an appropriate use of the scores yielded from this battery.

The test publishers claimed with Forms E and F that "the test assesses a pupil's norm-referenced performance in basic skills areas and provides objective-referenced information about a pupil and class attainment of specific objectives" (Airasian, 1989, p. 127). However, Airasian (1989) also cited that the authors were "properly cautious regarding interpretations of grade equivalent scores and provide examples of improper interpretations" (p. 127). This caution was absent in earlier editions of the technical manuals that were examined. In Wardrop's (1989) review of the assessment, stated that he was surprised that there was not more specific information regarding the use of test scores. Even though this researcher requested documents from CTB McGraw-Hill on more than one occasion, the documents that pertained to the specific recommended uses of the test scores was not provided.

In the 1993 manual, Macmillan/McGraw-Hill contended that "individual test results can help teachers plan specific learning strategies and activities" (p. 1-1). The manual goes on to identify the administrative decisions that are appropriate to make with the CAT. "Achievement test information, used in conjunction with other available information, can help educators make critical decisions related to placement of students into special programs, promotion, graduation, and attainment of prescribed competencies" (p. 1-1). Other appropriate uses identified by the publisher are class grouping, instructional program planning (that focuses on student need), curriculum analysis, needs assessment, program evaluation, and community relations (Macmillan/McGraw-Hill, 1993). The 1993 manual specifically stated, "A standardized achievement test is meant to sample the curriculum in the areas it measures. . . . Achievement tests must be kept in perspective. . . . they do not tell the whole story. Participation in classroom activities, classroom tests, homework assignments, and special projects can contribute to the evaluation of students' progress" (pp. 1-2 and 1-3.). The publisher stated, "No single test battery, no matter how comprehensive, can measure all of the desirable outcomes of an educational program" (p. 1-3).

Johnson (2005) contended that the results of the *TerraNova Second Edition*, otherwise known as the *California Achievement Test* and the *Comprehensive Test of Basic Skills*, could be used "to track progress over years and grades, to make decisions in a criterion-referenced manner about individual student's strengths and weaknesses, to plan additional instruction, and to report student's progress to parents" (p. 1034). However, it was impossible for this researcher to be able to analyze original sources. Despite several attempts to obtain copies of the guides and manuals that prescribe these uses, the publishing company did not provide those documents.

The Comprehensive Tests of Basic Skills

In the 1972 *Seventh Mental Measurements Yearbook,* Brown (1972) stated that the strength of the Comprehensive Test of Basic Skills was the "emphasis [by the publisher] placed on using the tests to plan, evaluate, and improve instruction, and to help individual student learn, rather than just to rank students" (p. 23). The scores that were provided "can be of great value in improving instruction" (Brown, 1972, p. 23).

Hopkins (1992) maintained, "Little information has been provided to assist user in interpreting and using test scores beyond the rich description of the content and skill measured by the test battery" (p. 8). Hopkins (1992) asserted that there was "little evidence. . . reported for specific uses of the scores" (p. 8). Hopkins' (1992) assessment of the lack of information was consistent with the evidence in this study. The publisher of the Comprehensive Test of Basic Skills (CTBS), CTB McGraw-Hill, did not respond with documents in the form of technical manuals or guides that addressed the specific uses of test scores from the CTBS.

Secondary sources, such as Monsaas (2001), argued that the "Teacher's Guide and a separate guide for interpreting test scores and using test results" were very "useful" (p. 1223). Monsaas (2001) also held that the scores were to be sued to look at individual and group status over time. She also asserted that the scores could be used to "provide information about the effectiveness of educational programs" (2001, p. 1223). Johnson (2005), in a review of the assessment, professed that the guides of the *TerraNova* were designed "'to provide achievement scores that are valid for most types of educational decision making'" (p. 1034). The publishers identified that the scores could be used to "track progress over years and grades, to make decisions . . . about individual student's strength and weaknesses, to plan additional instruction, and to report student's progress to parents (Johnson, 2005, p. 1034).

The National Assessment for Educational Progress

According to the National Center for Education Statistics (October 2005), there are two goals of the National Assessment for Educational Progress: "to compare student achievement is states and other jurisdictions and to track changes in achievement. . . over time in mathematics, reading, writing, science, and other content domains" for students in grades 4, 8, and 12. NAEP data are made available to those in educational research, and Congress has provided for "ongoing evaluation" of the assessment (National Center for Education Progress, October 2005). In fact, the National Center for Education Statistics (NCES) and the Educational Testing Service (ETS) provide workshops on data analysis to assist in these efforts (The NAEP Guide, 1999). Additionally, the NCES conducts seminars on the Use of the NAEP Database for Research and Policy Discussion that "stimulates interest in using NAEP data to address educational research questions, . . . and demonstrates the steps necessary for conducting accurate statistical analyses of NAEP data" (The NAEP Guide, 1999, p. 49).

According to Mosher (2004), the original method of score reporting for NAEP was based on the fact that the scores should be interpreted according to what students know and can do. However, even though the test itself had not changed much, the uses for the scores had changed. When the scores began to be reported as scale scores and achievement levels, the shift to Item Response Theory as espoused by norm-referenced tests was more pronounced. The NAEP was no longer a criterion-referenced assessment. This may have led to providing "ambiguous or misleading information" (p. 337). After 1989, however, the NAEP began to take into consideration the recommendations of the National Council of Teachers of Mathematics (NCTM) and adopted the language from NCTM that "students should develop 'mathematical power'. . . [and] apply their mathematical knowledge to the solution of real-world problems"

(Mosher, 2004, p. 338). Interestingly, Mosher noted, these are not the skills that schools teach students. So when the NAEP reported that students are below basic, the scores may have reported aptitude and not learning of what was being taught in American schools.

Francis Keppel, the Commissioner of Education for the United States in 1963, enlisted the help of Ralph Tyler to explore the possibility of an assessment that eventually became the NAEP. While Tyler thought to create an assessment that looked at student achievement over time, Keppel "wanted to have national data that would meet the intent of the legislation that created a Department of Education" (Lehmann, 2004, p. 28). And it was the opinion of many that federally funding such an assessment automatically put it in the "political arena" (Lehmann, 2004, p. 28).

Summary

This chapter examined what the test publishers' and test authors' intention is regarding the use of their tests' scores. Although some of the earliest guides promoted the use of the tests for the purposes of grouping and classifying students, most of the authors and publishers have felt from the beginning that their assessments should be used as just one source of information to make decisions about students. And as time has passed and accountability measures became the focus of standardized testing, most publishers increasingly and explicitly stated the inappropriate uses for the results of their tests. The inappropriate uses that are identified are as follows:

1. As the sole source of information for promotion, retention, or graduation,
2. To compare students, schools, school systems, and states, and
3. As the sole source of information to evaluate an entire educational program.

However, most of the publishers were comfortable with their tests being used to show progress over time and even to evaluate programs if the results were used as one piece of information in

the whole picture about a student or an educational program. The appropriate uses of test scores are as follows:

1. To track student progress over time across subtests,
2. To identify student strengths and weaknesses,
3. As an instructional tool to assist in grouping students for instruction, and
4. As a supplemental record to confirm or highlight discrepancies in the student's record,

Each publisher, however, was careful to note that the use of their tests should be done when the test user is well-informed about the types of information that could be gleaned from their respective test scores.

Chapter 6

IMPLICATIONS AND RECOMMENDATIONS

The foregoing review of the origins of standardized testing and the appropriate and inappropriate uses identified by testing experts and tests specifications yields several important findings. The chapter is centered around the three research questions stated at the outset: what were the social and political factors that led to the rise of the standardized testing movement?; What are the appropriate uses of standardized test scores as identified by educational measurement literature in the United States?; and What are the implications for contemporary educational policy?

What were the social and political factors that led to the rise of the standardized testing movement?

At the beginning of the twentieth century, the dramatic and sudden rise in the population and the numbers of children to be educated led to the widespread use of standardized tests. Between 1890 and 1917, the population of the United States doubled, primarily through immigration (Chapman, 1988). Yet another factor, according to Judd (1933) and Chapman (1988), in the boost of school attendance was the appearance of compulsory attendance laws. "School attendance was up markedly around this time, rising from 136 persons in 1,000 in 1904 to 152 persons in 1916" (Chapman, 1988, p. 43). By 1930, enrollment had soared from the 1900 figures of 700,000 to 4.8 million (Angus and Mirel, 1999).

When the percentage of enrollments as a proportion of the population that was aged 14-17 jumped from 5.6 percent in 1890 to 50.7 percent in 1930 (Angus and Mirel, 1999), administrators were forced to find ways to accommodate this varied student population. One

approach was to differentiate courses of study based on students aptitudes and aspirations. However, the question then arose: who follows what course of study? Testing seemed to be the answer to that question of how to sort students who came from such different backgrounds. The great diversity of students left educators to find an economically-feasible way to decide who should take what courses and how grouping should take place. Extensive forms of group testing that emerged from World War I seemed to solve this problem. The practice had already proven successful in Europe with the work of men like Binet. The most accepted American modification of Binet's work was by Terman, a professor at Stanford University. It was Terman and Thorndike who began to see Binet's work as so much more than Binet ever did. Terman and Thorndike were the ones who were "advocates of the widest possible use of IQ testing by American educators, so that students could be assessed, sorted, and taught in accordance with their capabilities" (Lemann, 2000, p.18).

During World War I, the use of the Army tests to determine which soldiers would have specific jobs and which would become officers became an efficient way of classifying soldiers. It was these examinations that led to the inception of the Army Alpha, which was the first large-scale use of intelligence tests (Dubois, 1970). The Army Beta, used for those who were illiterate or who did not speak English, also became popular for use in assessing the abilities of soldiers that were already in military service during World War I. Yerkes saw intelligence tests as a way for the Army to classify recruits, and he was joined in his opinion by Terman (Chapman, 1988).

Terman saw testing as a way to sort student in schools much the same way that the Army sorted and classified recruits. He viewed testing as a viable method of diagnosing individual students' strengths and weaknesses and for vocational guidance. This was accomplished through group testing that he thought would reveal intelligence. The numbers of tests being purchased

and administered to students rose dramatically during the 1930s. During this period, as well, the testing industry became big business.

During the 1940s and 1950s, the educational use of tests was largely diagnostic. The 1950s brought the refining of many tests, and this refining took the form of new editions of assessments that were used by professional educators in an effort to determine student strength and weaknesses. It was in the 1960s, however, that the use of test scores became highly politicized. The Civil Rights Movement and the need for equality was at the heart of the testing movement. Test scores from the SAT only proved that, no matter in which section of the country that the students were educated, black students from schools that were predominantly black did more poorly than blacks from integrated schools (Lemann, 2000). The Civil Rights Act of 1964 mandated equality in education among black and white students, and testing would tell policymakers if this equality was being achieved.

During the 1970s, the numbers of test revisions and the numbers of tests administered to students did not decrease. It was during this decade that SAT scores began to fall, and policymakers blamed the quality of schools and teaching for that decline. The 1983 publication of *A Nation at Risk* launched the accountability through testing movement. This report recommended standardized testing as a means of school reform. The follow-up to that report, *High School: A Report on Secondary Education in America*, recommended that a Student Achievement and Advisement Test, similar to the testing system in England, be implemented in Amercian schools (Haney, Madaus, and Lyons, 1993). Holding teachers, schools, and school systems accountable by ranking them according to test scores was seen as the way to force reform.

Throughout the 1990s, this trend continued and culminated with the legislation in 2001 known as *No Child Left Behind* that required schools, school systems, and states to test students at gateway years. *No Child Left Behind* (2001) requires that states enact a policy of assessment in grades 3, 5, and 8 each year in order to track progress and determine Adequate Yearly Progress. While the assessments are mandated to be criterion-referenced assessments, it is important to note that the principles of testing are the same whether the test in question is a norm-referenced test or a criterion-referenced test. Based on test scores and two other indicators (test participation rate and student attendance), the institutions are determined to make Adequate Yearly Progress or to be labeled as Needs Improvement. However, of the three indicators that are considered for schools to be deemed as having made Adequate Yearly Progress, only one, the test score, is an indicator of student learning. Test participation and student attendance do not measure student academic progress.

It is undeniable that educators, politicians, and parents desire improved educational practices in the United States. According to Haney and Madaus (1978), though, "The theme behind all of these ideas is that more systematic managment of education can improve results. The trouble with these proposals, most of them modeled after industrial practices, is that they overlook the greater complexity of schools and education" (p. 60). Popham (2004) pointed out, "This is a simple but important point—namely, that educational testing is far less precise than most parents (and numerous educators) think it is" (p. 54). Theodore Sizer was very clear in *Horace's Compromise* that tests "would not solve our educational problems" (Hayes, 2006, p. 73).

What are the appropriate uses of standardized test scores as identified by educational measurement literature in the United States?

There were three major appropriate uses of test scores that emerged from the study of the literature in the field of educational measurement in the United States.

1. Using test scores as *one* piece of data to make placement decisions about a student.
2. Using test scores as *one* piece of data to identify the strengths and weaknesses (particularly wide discrepancies) of a student.
3. Using test scores as *one* piece of data to evaluate an educational program, including the effectiveness of teachers, schools, school systems, or states.

These appropriate uses pervaded the literature throughout the twentieth century and continue to be appropriate in the twenty-first century. There were those scholars in the field who may have disagreed about specific points of the testing debate, but they all espoused the idea that these uses are appropriate.

Testing experts considered it acceptable to use test scores as a part of a larger set of data that was analyzed to make decisions about students. The rejection of using a single test score from a single test administration pervaded in the literature of educational measurement. Binet and Simon (1916), for example, declared, "Obviously it rests upon the principle that a particular test isolated from the rest is of little value, that it is open to errors of every sort, especially if it is rapid and is applied to school children. . . " (p. 329). The practice of using one test score to make important decisions about students was considered by Binet to be an unacceptable use of the otherwise helpful practice of using tests as one source of information about students.

Binet was only the first in a series of renowned scholars in the field of educational measurement to assert that a single test score did not provide sufficient information to measure a student's intelligence or academic progress. In the 1920s, Pressey and Pressey (1922) advocated for using tests in conjunction with other pieces of information about a student to make such important decisions such as those regarding promotion and retention. The opinion that test scores could be valuable but were not infallible persisted into the 1940s. Tyler (1944) proposed that "the greatest advances can now be made in admission and placement by the use of examinations, although we recognize the importance of stimulating schools and colleges to maintain more adequate records which include observations, samples of work, and other evidence of student abilities, interests, and accomplishment" (p. 11). Tyler (1944) acknowledged that other evidence gathered by the teacher is just as valuable in making such decisions about students.

Cronbach (1949) and others also asserted that it was never fitting to use just a single test score on a single assessment to make a decision about a student as important as retention, promotion, or graduation. This view prevailed through the 1960s during the development of the National Assessment for Educational Progress and into the 1980s. Chapman (1988), for example, reported, "Regarding the use of tests, one implication is that educators must use a variety of measures to assess talent and provide instruction tailored to individual needs" (p. 177). This wide array of measures included teacher observations, grades, and multiple assessment data including ongoing evaluations of student performance of both a formal and an informal nature. Instead of using single test scores to make high-stakes decisions, Madaus (1988) stated, "Testing programs should, in my view, be seen as an ancillary tool of curriculum and instruction,--albeit, a very necessary, useful, and important one—and nothing else" (pp. 84-85). As a supplement to

the student's record, a test score can be a useful piece of information that helps teachers track student progress over time.

Despite this largely recognized testing principle, Chapman (1988) observed that "tests are still misused where they are relied upon exclusively for making judgments about students (p. 177). This includes not only decisions about retention, promotion, or graduation—those decisions that are considered high-stakes—but also those regarding the strengths and weaknesses of students. Terman (1923) took the admonition to use the test score in conjunction with other pieces of information about the student a step further. He called for the test score to "be taken as the point of departure for further study of the pupil" (p. 25). For example, in the event that a student's test score did not match what the teacher knew of the other attributes of the student, then more study (perhaps further testing) and the gathering of additional data were necessary. Clearly, the use of a single test score to make high-stakes decisions was not supported. Rather any discrepancy between scores and other information was seen as an indicator that the student may have needed additional assistance. This use encouraged teachers, counselors, and administrators to look at multiple source of information about a student before making important decisions. This was a suitable way to use test scores according to testing experts.

In addition to considering test scores as one piece of data in making decisions about students, test experts emphasized that test scores should be only one piece of the data used to make decisions about an educational program, including the quality or effectiveness of teachers. The Presseys (1922) cautioned not only against using a single test score to make critical decisions about students, but alsoagainst using test scores to determine the quality of teachers or schools. Just as students could not be judged by one test score on one day, neither should a

teacher or the educational program in a particular school, school system, or state be appraised based on a sole performance on a standardized test.

Ansley (2000) asserted that local boards of education should be interested in their schools' scores on a battery of tests as "a single piece of a fairly large and complex puzzle of educational achievement" (p. 279). Ansley (2000) was also critical of using standardized test scores to determine the effectiveness of individual teachers. "This clearly represents a gross misuse of test scores" (Ansley, 2000, p. 279). It was deemed impossible to determine the overall value in a particular classroom or program based on the performance of one group of students on one assessment. Student demographics, teacher experience, and other factors that influence student achievement must be contemplated before making judgments about teachers and schools.

What are the implications for contemporary educational policy?

It is a well-established principle of the scholars and experts in the field of educational measurement that a single test score should not be used to make important decisions. Such decisions include promotion, retention, graduation, teacher and school effectiveness, and program quality. Although this principle was established correctly at the advent of standardized testing and has been reaffirmed by testing experts since then, the public policy that educators are required to follow in the twenty-first century demands that they do the opposite of what has been established as the suitable uses of tests. Educators are expected to retain students if they do not make a specific score on a certain test. Educators are even required to prevent students from graduating from high school for not passing a test. Educators are judged as incompetent and in need of improvement if students do not reach a certain annual measurable objective as determined by a single test score.

In terms of politics, standardized testing has had the symbolic function of assuring the public that "something is being done about poor educational performance" (Mazzeo, 2001, p. 372). Negative public perception of public schools has been fueled the media's coverage of politicians' assertions that they are doing something about the state of our schools by mandating accountability. Madaus (1988) argued, "Policymakers are well aware of the high symbolic value tests and test results can have in creating an image of progress or reform. By mandating a test, policymakers are seen to be addressing critical reform issues forcefully, in a way the public understands" (p. 89) However, what many parents and the public in general fail to realize is that when politicians hold students, teachers, schools, school systems, and states accountable by mandating standardized tests, they are using faulty measures. Principles of educational measurement are clear that using a single test score to make high stakes decisions about students is not sound practice, but using a single score is exactly what legislation such as No Child Left Behind and the A+ Education Reform Act do. In reality, standardized tests can provide teachers with useful data about students so that instruction can be tailored to meet the needs of individual students. However, the abuse and misuse of the scores to make decisions about retention, promotion, graduation, and funding for schools is what is in direct contradiction to established principles and even contradicts what the test publishers themselves outline as the appropriate uses for their scores.

In fact, many politicians and policymakers have convinced parents that holding schools accountable for scores on standardized tests will bring about much-needed reform in schools. They have assured the public that if schools are judged to be ineffective based on those test scores, then they will change their practice so that students will perform well on the tests. According to Lemann (1999), aptitude measures "have the least reforming effect" (Lemann,

1999, p. 7). In fact, the focus on using a single score for accountability has created substandard schooling in the United States. Research has found that high-stakes tests compel teachers to teach only what is on the standardized test and to teach that material in a manner that mimics the format of the test. As Madaus (1988) pointed out, "As test scores rise over time, policymakers point to the wisdom of their action and the general public's confidence in the schools is restored. However, the real possibility that the testing program may not be a cure for the underlying problem, and the reality of the power of such programs to distort the educational process must eventually be faced" (p. 89).

One implication of the widespread use of standardized test scores is that teachers begin to use test preparation materials in a way that is mechanical and fails to engage children in their learning (Hill, 2001). Madaus (1988) stated, "Teachers see the kind of intellectual activity required by previous test questions and prepare the students to meet these demands" (p. 93). Many teachers and administrators are forced to follow a program of test preparation just to be sure that their students are prepared to take and pass the assessments. The kinds of activities in which many schools engage are those that do not promote critical thinking skills and rigor because these are not the kinds of questions that are asked on a standardized achievement test (Madaus, 1988). Instead of having a reforming effect, instruction becomes "dominated by tasks that resemble tests" (Shepard, 1991, p. 233). Students are subjected to worksheets and skill and drill activities that will help them score high on tests but have detrimental effects such as destroying their love of reading (Neill and Medina, 1989). While teachers spend so much time on the drilling of skills that is required for students to do well on standardized tests, "Among the instructional casualties are the higher-order thinking skills" (Neill and Medina, 1989, p. 694). This is probably the most critical error that educators make when a high score on a standardized

test becomes the focus of the curriculum. The failure to teach young students higher order thinking skills is something that impacts the quality of their learning for the rest of their educational careers. The impact is felt beyond their schooling years and into their college and work lives as they are faced with problems and situations that they cannot successfully think their way through. Furthermore, test results are further skewed because the perceived improvement on the test is not the result of improved student achievement but of students learning what is expected to be on the test (Madaus, 1988). Based on those test results, if a school does not make Adequate Yearly Progress, then teachers and principals wonder just what they could have done differently in terms of preparing students for the test. In reality, there are many other indicators that demonstrate the progress of students that give a very different picture than what is represented by a single test score published in the newspaper.

One further effect to which politicians and policy makers give little thought is the fact that legislation like No Child Left Behind and the A+ Education Reform Act sends a very clear message to teachers: "the public and policy makers mistrust the teacher's judgment" (Madaus, 1985, p. 615. Teachers are seen by many parents and politicians as people who are not the experts, not the ones who see the students' performance on a daily basis, and who are quite incapable of deciding if a student is making progress without a single test score to confirm their opinion. Even if a teacher has multiple pieces of evidence to demonstrate the gains of an individual student or a whole class, that evidence is discounted if a single test score gives different information. Madaus (1988) also claimed, "In a high-stakes testing program teachers cannot ignore results or treat them as occasional experiences, or interpret them in light of their hidden knowledge. The results leave no room for teacher input into the decision" (p. 101).

Therefore, four recommendations for educational policy emerged from this study, and they are as follows:

1. Scores on standardized tests should be used as one source of data among several used to make high-stakes decisions about students.

2. Scores on standardized tests should be used as one source of data among several used to evaluate schools, teachers, and educational programs.

3. The content addressed by items on standardized tests should not be used as the sole source of curriculum.

4. Scores on standardized tests should not be used to rank schools or to foster competition between schools and states.

Implications for Further Research

The following possible lines of research emerge from this study. Additional research on several of the seven assessments included in this study would be useful. Some of the primary source documents for the Stanford Achievement Tests, the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the California Achievement Test, and the Comprehensive Test of Basic Skills were not available. Whereas this study in places had to rely on secondary sources of information, access to the primary sources would complement this study. Greater cooperation from the owners of out-of-print documents would be essential.

Supplementary research on the effects of Measurement-Driven Instruction would be useful. Since the largest discussions of Measurement-Driven Instruction took place in the 1980s, very little has been written of the concept. With the passage of No Child Left Behind, assessments continue to drive instruction in schools. Because the idea stemmed from the belief that what was tested on particular assessments determined what teachers would teach and what

students would learn, it is particularly true because the tests are being used to rank schools. Madaus (1988) found that high-stakes tests narrowed the curriculum. With the emphasis on assessments as high-stakes instruments greater than ever, additional research into the effect on curriculum would be valuable.

Further research could also take the form of studying the function that NAEP can serve in terms of such assessments. The mandate in No Child Left Behind that all states must participate in the NAEP at the local level when the school systems are chosen for participation has actually created a unique opportunity for NAEP to take on this role. According to Vinovskis (1998), by itself, NAEP is not sufficient to determine student achievement or to rank students, schools, and even teachers. Instead, that was not its original function. However, more focus on the development of appropriate, standards-based performance tasks may move NAEP forward as a tool to assist parents, educators, and policymakers in determining the true educational progress that American students are making. Slavin (1997) in *Educational Reseracher* quoted Vinovskis as follows:

> For decades, policy makers have complained that the federal research and development enterprise has had too little impact on the practice of education. . . . The limited direct influence of federal R & D compared to that of, say, research in medicine, physics, and chemistry can certainly be ascribed in part to the far more limited federal investment in educational R & D. . . . (p. 67).

If the NAEP continued to align itself to the standards of multiple subject areas as it has with the National Council of Teachers of Mathematics, then it would lend itself to research on how standards and performance assessments impact student progress over time.

Further research could be conducted on performance assessments as one of the most promising answers to school reform. The research project could be conducted with a school system or consortium of school systems that represent a variety of demographic groups to be released from the mandates of No Child Left Behind. These schools and school systems could then create a culture of change that reflects standards-based instruction and performance assessments of that instruction. Inside education, reform has taken form of developing standards that will improve upon an educational system that taught relatively the same thing for the previous 100 years (Firestone and Schorr, 2004). As Lewis (2000) asserted, such instruction "is essential to closing the gap" (p. 103), but not enough schools, systems, and states were implementing standards-based classroom instruction that leads to performance assessments. Many states such as Kentucky, Maryland and Vermont have implemented the use of performance assessments, and additional longitudinal research into the progress that students are making in such states as compared to students in other states might provide insight as to the effectiveness of such assessment programs. These programs often include portfolio assessments, a set of performance tasks, and student self-assessment. Lewis (2000) also maintained that this movement may be the key to helping educators eliminate the disparity that plagues low-achieving students (especially minority students). Significantly, the use of performance assessments is consistent with calls from testing experts for multiple sources of data about student learning.

Performance assessments are an important beginning step to encourage deeper teaching of subjects, and the change must be pervasive and not limited to those subjects that are tested in particular grade levels. Firestone and Schorr (2004) also emphasized that the most effective accountability measures are those that keep the external stakes high enough to reduce the

negative consequences and at the same time build the internal accountability that comes with changing the culture of a school or school system. Teachers who feel both pressure and support in a balanced way tend to change instructional practices in a positive way that truly results in meaningful reform for students. In this culture shift, the school or system determines what is important for their students to learn, sets up the environment for them to learn it, and then internally polices themselves when it does not happen. This kind of culture shift requires support at all levels in the school or system (Firestone and Schorr, 2004). In order to accomplish this goal, then it is necessary for the public, especially parents, to be given the proper information to understand that this is a different kind of evaluation than is customary in the United States. Educating the public on performance-based assessments and standards-based instruction and the limitations of single score assessments are important facets of meaningful reform (Firestone and Schorr, 2004).

REFERENCES

Abbott, S. (1997). *Standardized testing*. Westminster, CA: Teacher Created Materials.

Airasian, P. W. (1989). Review of the California Achievement Tests. In J. C. Conoley and J. J. Kramer's (Eds.) *The tenth mental measurement yearbook*. (pp. 126-128). Lincoln, NE: University of Nebraska Press.

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association (2007). *The code of fair testing practices in education.* Retrieved July 15, 2007 from http://www.apa.org/science/fairtestcode.html#a.

Angoff, W. H. and Dyer, H. S. (1971). The admissions testing program. In W. H. Angoff's (Ed.), *The College Board admissions testing program: a technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests*. (pp. 1-13). New York, NY: College Entrance Examination Board.

Angus, D. L. and Mirel, J. E. (1999). *The failed promise of the American high school, 1890-1995*. New York, NY: Teachers College Press.

Ansley, Tim (2000). The role of standardized achievement tests in grades k-12. In A. D. Trice's (Ed.) *Handbook of Classroom Assessment*. New York, NY: Longman.

Apple, M. W. (2000). Standards, markets, and curriculum. In B. M. Franklin's (Ed.). *Curriculum and consequence: Herbert M. Kliebard and the promise of schooling*. (pp. 55-74). New York, NY: Teachers College Press.

Bagley, W. C. (1925). *Determinism in education*. Baltimore, MD: Warwick and York, Inc.

Bauernfeind, R. H. (1978). *Building a school testing program: uses and misuses of standardized tests*. Bensenville, IL: Scholastic Testing Service.

Berk, R. A. (1998). Review of the Stanford Achievement Test. In J. C. Impara and B. S. Plake's (Eds.) *The thirteenth mental measurements yearbook*. (pp. 925-928). Lincoln, NE: University of Nebraska Press.

Binet, A. and Simon, T. (1916). *The development of intelligence in children.* Baltimore, MD: Williams and Wilkins Company.

Bowles, Frank (1949). Personal Memorandum to Principals and Headmasters, November 22, 1949.

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

Bracey, G. W. (1987). The muddles of measurement-driven instruction. *Phi Delta Kappan*, 68:9, 688-689.

Bracey, G. W. (1998*). Put to the test: an educator's and consumer's guide to standardized testing*. Bloomington, IN: Phi Delta Kappa International.

Brigham, C. C. (1933). Report to the executive committee of the College Entrance Examination Board. New York, NY: College Entrance Examination Board.

Brown, F. G. (1972). Review of the Comprehensive Tests of Basic Skills. In O. K. Buros' (Ed.), *The seventh mental measurements yearbook.* (pp. 21-23). Highland Park, NJ: The Gryphon Press.

Bryan, M. M. (1965). Review of the Stanford Achievement Tests. In O. K. Buros' (Ed.), *The sixth mental measurements yearbook*. (pp. 110-124). Highland Park, NJ: The Gryphon Press.xx.

Bryan, M. M. (1978). Review of the California Achievement Tests. In O. K. Buros' (Ed.), *The eighth mental measurements yearbook*. (pp. 35-37). Highland Park, NJ: The Gryphon Press.

California Test Bureau (1951). *California Test Bureau manual: California Achievement Tests Complete* Battery. Los Angeles, CA: Calfornia Test Bureau.

California Test Bureau (1957). *Technical report on the California Achievement Tests.* Los Angeles, CA: California Test Bureau.

California Test Bureau (1961). *Questions and answers on the California Achievement Tests.* Monterey, CA: California Test Bureau.

Calvin, A. (2000). Use of standardized tests in admissions in postsecondary institutions of higher education. *Psychology, Public Policy, and Law*, 6:1.

Cameron, R. G. (1989). *The common yardstick: a case for the SAT*. New York, NY: College Entrance Examination Board.

Campbell, H. W. (1957). Candidates and tests, 1957-58. New York, NY: College Entrance Examination Board.

Cannell, J. J. (1989). *The "Lake Wobegon" report: how public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends of Education.

Carney, R. N. (2005). Review of the Stanford Achievement Test. In R. A. Spies and B. S. Plake's (Eds.), *The sixteenth mental measurements yearbook*. Retrieved on October 1, 2007 from http://search.ebscohost.com/login.aspx?direct=true&db=loh&AN16013155&site=ehost-live

Cattell, R. B. and Moodie, W. (1936). *A guide to mental testing for psychological clinics, schools, and industrial psychologists.* London: University of London Press.

Chapman, P. D. (1988). *Schools as sorters : Lewis M. Terman, applied psychology, and the intelligence testing movement, 1890-1930*. New York, NY: New York University Press.

Chase, C. I. and Ludlow, H. G. (Eds.). (1966). *Readings in educational and psychological measurement.* Atlanta, GA: Houghton Mifflin Company.

Civil Rights Act of 1964. Public Law No. 88-352, 78 Stat. 241 (1964).

Clifford, G. J. (1984). *Edward L. Thorndike: the sane positivist*. Middletown, Connecticut: Wesleyan University Press.

College Entrance Examination Board (1956). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1957). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1958). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1960). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1961). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1962). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1963). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1964). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1965). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1966). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1967). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1968). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1969). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1970). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1971). *A description of the College Board Scholastic Aptitude Test*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1960). *SAT: a guide for counselors.* Princeton, NJ: College Entrance Examination Board.

College Entrance Examination Board (1964). *SAT: a guide for counselors and admissions officers*. Princeton, NJ: College Entrance Examination Board.

College Entrance Examination Board (1965). *SAT: a guide for counselors and admissions officers*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1926). *Scholastic Aptitude Tests: a manual for the use of schools*. New York, NY: College Entrance Examination Board.

College Entrance Examination Board (1949). *Supplement to the 1949 handbook—changes in terms of admission to the member colleges for 1950*. New York, NY: College Entrance Examination Board.

CTB/McGraw-Hill (n. d.). *Technical report: California Achievment Tests forms e and f levels 10-20.* Monterey, CA: CTB/McGraw-Hill.

Cremin, L. A. (1976). *Public education*. New York, NY: Basic Books, Inc.

Cronbach, L. J. (1949). *Essentials of psychological testing: third edition*. New York, NY: Harper and Row.

Cureton, E. E. (1951). Validity. In E. F. Lindquist's (Ed.), *Educational measurement.* (pp. 621-694). Washington, DC: American Council on Education.

Davis, O. L., Jr. (1991). Historical inquiry: telling real stories. In E. C. Short (Ed.), *Forms of curriculum inquiry* (pp. 77-88). Albany, NY: SUNY Press.

Deighton, L. C., ed. (1971). Encyclopedia of education. Vol 4.

Deming, W. C. (1923). Civil service and duties of commission explained. *Congressional Digest*. 2: 7. 198-199.

Dewey, J. (1904). *The educational situation*. Chicago, IL: The University of Chicago Press.

Donlon, T. F., editor. (1984). *The College Board technical handbook for the scholastic aptitude test and achievement tests*. New York, NY: College Entrance Examination Board.

Donlon, T. F. and Angoff, W. H. (1984). The Scholastic Aptitude Test. In T. F. Donlon's (Ed.) *The College Board technical handbook for the scholastic aptitude test and achievement tests*. New York, NY: College Entrance Examination Board.

Dubois, P. H. (1970). *A history of psychological testing.* Boston: Allyn and Bacon.

Dyer, H. S. and King, R. G. (1954). *College Board scores: their use and interpretation. No. 2.* New York, NY: College Entrance Examination Board.

Ebel, R. L. (1977). *The uses of standardized testing.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Educational Testing Service (2007). SAT Timeline.

Epstein, J. (1996). The National Assessment for Educational Progress and educational policy making. In P. S. Hlebowitsh and W. W. Wraga's (Eds.) *Annual review of research for school leaders*. (pp. 22-43). New York, NY: Scholastic.

Findley, W. G. (1953). Review of the Metropolitan Achievement Tests. In O. K. Buros' (Ed.), *The fourth mental measurements yearbook*. Highland Park, NJ: The Gryphon Press.

Findley, W. G. (1965). Review of the Metropolitan Achievement Tests. In O. K. Buros' (Ed.), *The sixth mental measurements yearbook*. Highland Park, NJ: The Gryphon Press.

Findley, W. G. (1978). Review of the Comprehensive Tests of Basic Skills. In O. K. Buros' (Ed.), *The eighth mental measurements yearbook*. (pp. 40-43). Highland Park, NJ: The Gryphon Press.

Finley, C. J. (1995). Review of the Metropolitan Achievement Test. In J. C. Conoley and J. C. Impara's *The twelfth mental measurements yearbook*. (pp. 603-606). Lincoln, NE: University of Nebraska Press.

Firestone, W. A., and Schorr, R. Y. (2004). Introduction. In W. A. Firestone, R. Y. Schorr, and L. F. Monfil's (Eds.) *The ambiguity of teaching to the test: standards, assessment, and educational reform*. Mahweh, NJ: Lawrence Erlbaum Associates.

Fraenkel, J. R. and Wallen, N. E. (1996). *How to design and evaluate research in education.* St. Louis, MO: McGraw-Hill.

Gardner, H. H. (1923). Civil service procedure. *Congressional Digest*. 2:7, 199.

Georgia Department of Education. (2002, August). *Implementing no child left behind*. Retrieved February 14, 2003, from http://www.doe.k12.ga.us/_documents/support/plan/nclb_part_2.pdf .

Hambleton, R. K. (1995). Review of the Metropolitan Achievement Test. In J. C. Conoley and J. C. Impara's *The twelfth mental measurements yearbook*. (pp. 606-610). Lincoln, NE: University of Nebraska Press.

Haney, W. M. and Madaus, G. F. (1978). Making sense of the competency testing movement. In P. W. Airasian, G. F. Madaus, and J. J. Pedulla's (Eds.) *Minimal competency testing*. (pp. 49-72). Englewood Cliffs, NJ: Educational Technology Publications.

Haney, W. M., Madaus, G. F., and Lyons, R. (1993). *The fractured marketplace for standardized testing.* Boston, MA: Kluwer-Academic Publishers.

Harcourt Assessment (2004). *Stanford Achievment Test Series, ninth edition, technical data report*. San Antonio, TX: Harcourt Assessment, Incorporated.

Harcourt Assessment. (2006). *History*. Retrieved January 12, 2007, from https://harcourtassessment.com/haiweb/Cultures/en-US/Harcourt/AboutUs/CompanyInformation/History .

Harcourt Brace and Company (1994). *A history of Harcourt Brace and Company: 75 years of publishing excellence*. Orlando, FL: Harcourt Brace and Company.

Harcourt Brace Educational Measurement (1997). *Stanford Achievment Test Series, ninth edition: technical data report*. Atlanta, GA: Harcourt Brace and Company.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, advanced ½: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, intermediate 1/2/3: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, primary 1: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, primary 2: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, primary 3: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, sesat ½: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harcourt Educational Measurement (2003). *Stanford Achievement Test Series, tenth edition, task 1/2/3: guide for classroom planning*. San Antonio, TX: Harcourt Educational Measurement.

Harris, L. A. (1978). Review of the Iowa Test of Basic Skills. In O. K. Buros's (Ed.) *The eighth mental measurements yearbook*. (pp. 55-57). Highland Park, NJ: Gryphon Press.

Hawkes, H. E., Lindquist, E. F., and Mann, C. R. eds. (1936). *The construction and use of achievement examinations: a manual for secondary school teachers*. Atlanta, GA: Houghton Mifflin Company.

Hayes, W. (2006). *The progressive movement: is it still a factor in today's schools?* New York, NY: Rowman and Littlefield Education.

Hill, Clifford (2001). The pitfalls of annual testing. *Christian Science Monitor*, 94.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., Bray, G. B., Lewis, J. C., Qualls, A. L., Mengeling, M. A., and Shannon, G. P. (1993). *The Iowa Tests guide to research and development: forms a and b levels 5-14.* Itasca, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Dunbar, S. B., Frisbie, D. A., Oberley, Cantor, N. K., Bray, G. B., Lewis, J. C., and Qualls-Payne, A. L. (1993*). The Iowa Tests interpretive guide for teachers and counselors: forms k and l levels 5-8*. Chicago, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Dunbar, S. B., Frisbie, D. A., Oberley, Cantor, N. K., Bray, G. B., Lewis, J. C., and Qualls-Payne, A. L. (1994). *The Iowa Tests interpretive guide for school administrators: forms k and l levels 5-14*. Chicago, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., and Qualls, A. L. (1996). *Iowa Test of Basic Skills interpretive guide for school administrators: form M levels 5-14 complete and survey*. Itasca, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., and Qualls, A. L. (1996). *Iowa Test of Basic Skills interpretive guide for*

teachers and counselors: form M levels 9-14 complete and survey. Itasca, IL: Riverside Publishing.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., Bray, G. B., Lewis, J. C., and Qualls, A. L. (2003). *The Iowa Tests interpretive guide for school administrators: forms a and b levels 5-14*. Itasca, IL: Riverside Publishing.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., Bray, G. B., Lewis, J. C., and Qualls, A. L. (2003). *The Iowa Tests interpretive guide for teachers and counselors: forms a and b levels 5-8*. Itasca, IL: Riverside Publishing.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., Bray, G. B., Lewis, J. C., and Qualls, A. L. (2003). *The Iowa Tests interpretive guide for teachers and counselors: forms a and b levels 9-14*. Itasca, IL: Riverside Publishing.

Hopkins, K. D. (1992). Review of the Comprehensive Test of Basic Skills, fourth edition. Retrieved October 1, 2007 from http://search.ebscohost.com/login.aspx?direct=true&db=loh&AN=11010983&site=ehost-live

Horn, E. (1926). Discussion of the general statement. In G. M. Whipple's (Ed.). *The twenty-sixth yearbook of the National Society for the Study of Education: the foundations and technique of curriculum-construction*. (pp. 99-112). Bloomington, IL: Public School Pubishing Company.

Johnson, R. L. (2005). Review of the TerraNova, the second edition. In R. A. Spies and B. S. Plake's (Eds.) *The sixteenth mental measurements yearbook*. (pp. 1030-1035). Lincoln, NE: University of Nebraska Press.

Joncich, G. (1968). *The sane positivist: a biography of Edward L. Thorndike*. Middletown, CT: Wesleyan University Press.

Jones, L. V. (1996). A history of the National Assessment for Educational Progress and some questions about its future. *Educational researcher*, 25 (7), 15-22.

Jones, L. V. and Olkin, I., eds. (2004). *The nation's report card: evolution and perspectives*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Judd, C. H. (1916). *Measuring the work of the public school*. Cleveland, OH: The Survey Committee of the Cleveland Foundation.

Judd, C. H. (1927). *Psychology of Secondary Education*. Atlanta, GA: Ginn and Company.

Judd, C. H. (1928). *The unique character of American secondary education.* Cambridge, MA: Harvard University Press.

Judd, C. H. (1933). *Problems of education in the United States*. New York, NY: McGraw-Hill and Company.

Judd, C. H. (1934). *Education and social progress*. New York, NY: Harcourt, Brace, and Company.

Kaestle, C. F. (1988). Recent methodological developments in the history of Amercian education. In R. M. Jaeger (Ed.), *Complementary methods for research in education.* (pp. 61-78). Washington, DC: American Educational Research Association.

Kliebard, H. M. (2002). *Changing course: American curriculum reform in the 20th century*. New York, NY: Teachers College Press.

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

Kreitzer, A. and Madaus, G. F. (1995). The test-driven curriculum. In D. Tanner and J. W. Keefe's (Eds.) *Curriculum issues and the new century*. (pp. 23-37). Reston, VA: National Association of Secondary School Principals.

Lagemann, E. C. (2000). *An elusive science: the troubling history of education research.* Chicago, IL: The University of Chicago Press.

Lehmann, I. J. (2004). The genesis of NAEP. In L. V. Jones and I. Olkin (Eds.) *The nation's report card: evolution and perspectives* (pp. 25-76). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Lemann, N. (1999). Behind the SAT. *Newsweek*, 134:10.

Lemann, N. (2000). *The big test: the secret history of the American meritocracy*. New York, NY: Farrar, Straus, and Giroux.

Lewis, A. C. (2000). The notorious g-a-p. *Phi Delta Kappan*, 82, 103-104.

Lindquist, E. F. (1944). Nationally coordinated regional testing programs in high schools. In American Council on Education Studies' *New directions for measurement and guidance*. (pp. 87-103). Washington, DC: American Council on Education.

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* Washington, DC: American Council on Education.

Linn, R. L. (1985). Review of the Comprehensive Test of Basic Skills. In J. V. Mitchell's (Ed.) *The ninth mental measurement yearbook*. (pp. 382-386). Lincoln, NE: University of Nebraska Press.

Lukin, L. E. (2005). Review of the Metropolitan Achievement Test. In R. A. Spies and B. S. Plake's (Eds.) *The sixteenth mental measurements yearbook*. Retrieved on October 1, 2007 from

http://search.ebscohost.com/login.aspx?direct=true&db=loh&AN=16012695&site=ehost-live

Macmillan/McGraw-Hill. (1993). *CAT/5 Technical Manual*. Monterey, CA: Macmillan/McGraw-Hill School Publishing.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66:9, 611-617.

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

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner's (Ed.) *Critical issues in curriculum: eighty-seventh yearbook of the National Society for the Study of Education*. Chicago, IL: University of Chicago Press.

Madaus, G. F. and Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson's (Ed.) *Handbook of research on curriculum*. New York, NY: Macmillan Publishing Company.

Marius, R. (1999). *A short guide to writing about history*. New York, NY: Longman.

Mazzeo, C. (2001). Frameworks of state: assessment policy in historical perspective. *Teachers College Record*, 103:3, 367-397.

Merwin, J. C. (1965). Review of the California Achievement Tests. In O. K. Buros' (Ed.) *The sixth mental measurements yearbook*. Highland Park, NJ: The Gryphon Press.

Minton, H. L. (1988). *Lewis M. Terman: pioneer in psychological testing.* New York, NY: University Press.

Monsaas, J. A. (2001). Review of the TerraNova. In B. S. Plake and J. C. Impara's (Eds.) *The fourteenth mental measurements yearbook*. (pp. 1223-1226). Lincoln, NE: University of Nebraska Press.

Mort, P. R. and Gates, A. I. (1932). *The acceptable uses of achievement tests*. New York, NY: Teachers College Bureau of Publications.

Mosher, F. A. (2004). What NAEP really could do. In L. V. Jones and I. Olkin (Eds.) *The nation's report card: evolution and perspectives.* (pp. 329-340). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Mosier, C. I. (1966). A critical examination of the concepts of face validity. In C. I. Chase and H. G. Ludlow (Eds.), *Readings in educational and psychological measurement*. (pp. 103-113). Atlanta, GA: Houghton Mifflin Company.

National Center for Education Statistics. (1999). *The NAEP* Guide. United States Department of Education: Author.

National Center for Education Statistics. (2003, April). *The History of NAEP* Partners. Retrieved October 7, 2006, from http://nces.ed.gov/nationsreportcard/contracts/history.asp

National Center for Education Statistics. (2005, August). *NAEP and No Child Left Behind*. Retrieved October 7, 2006, from http://nces.ed.gov/nationsreportcard/nclb.asp

National Center for Education Statistics. (2005, October). *Policymakers' Frequently Asked Questions*. Retrieved October 7, 2006, from http://nces.ed.gov.nationsreportcard/policymakers/faqs.asp

National Center for Education Statistics. (2006, June). *NAEP Overview*. Retrieved October 7, 2006, from http://nces.ed.gov/nationsreportcard/about/

National Center for Education Statistics. (2006, August). *Chronology of National Assessment of Educational Progress (NAEP) Assessments from 1969 to 2005*. Retrieved October 7, 2006, from http://nces.ed.gov/nationsreportcard/about/assesshistory.asp

Neill, D. M., and Medina, N. J. (1989). Standardized testing: harmful to educational health. *Phi Delta Kappan*, 70: 9, 688-702.

Nitko, A. J. (2004). Review of the California Achievement Tests, fifth edition. *Mental Measurements Yearbook*, 13. Retrieved September 17, 2007, from the EBSCOhost database.

Nitko, A. J. (1978). Review of the Comprehensive Tests of Basic Skills. In O. K. Buros' (Ed.), *The eighth mental measurements yearbook*. (pp. 43-45). Highland Park, NJ: The Gryphon Press.

No Child Left Behind Act of 2001, Public Law No. 107-110, 115 Stat. 1425 (2001).

Noble, A. J. and Smith, M. L. (1994). *Measurement-driven reform: research on policy, practice, repercussion*. Los Angeles, CA: National Center for Research on Evaluation.

Notestein, A. C. editor (1948). *Annual handbook 1948: terms of admission to the colleges of the College Entrance Examination Board*. Princeton, NJ: College Entrance Examination Board.

Null, J. W. (2003). *A disciplined progressive educator: the life and career of William Chandler Bagley*. New York, NY: Peter Lang.

Olsen, M. (1958). *Candidates and tests, 1958-59*. New York, NY: College Entrance Examination Board.

Olsen, M. (1959). *Candidates and tests, 1959-60*. New York, NY: College Entrance Examination Board.

Passow, A. H. (1977). *American secondary education: the Conant influence*. Reston, VA: National Association of Secondary School Principals.

Passow, A. H.. (1978). Review of the Metropolitan Achievement Test. In O. K. Buros's (Ed.) *The eighth mental measurements yearbook*. (pp. 102-105). Lincoln, NE: University of Nebraska Press.

Peterson, J. J. (1983). *The Iowa testing programs: the first fifty years*. Iowa City, IA: University of Iowa Press.

Popham, W. J. (Ed.). (1971). *Criterion-reference measurement: an introduction*. Englewood Cliffs, NJ: Educational Technology Publications.

Popham, W. J. (Ed.). (1974). *Evaluation in education*. Berkeley, CA: McCutchan Publishing Corporation.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.

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

Popham, W. J. (1987). Muddle-minded emotionalism. *Phi Delta Kappan*, 68:9, 687-688.

Popham, W. J. (1988). *Educational evaluation*. Englewood Cliffs, NJ: Prentice Hall.

Popham, W. J. (1990). *Modern educational measurement: a practitioner's perspective*. Englewood Cliffs, NJ: Prentice Hall.

Popham, W. J. (2000). *Testing! testing!*. Boston, MA: Allyn and Bacon.

Popham, W. J. (2001). *The truth about testing: an educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Pressey, S. L. and Pressey, L. C. (1922). *Introduction to the use of standard tests: a brief manual in the use of tests of both ability and achievement in the school subjects*. Yonkers-on-Hudson, New York, NY: World Book Company.

The Psychological Corporation (1985). *Stanford Achievement Test Series, eighth edition: technical data report*. Atlanta, GA: Harcourt Brace Jovanovich, Incorporated.

The Psychological Corporation (1988). *Metropolitan Achievement Tests sixth edition survey battery technical manual.* Atlanta, GA: Harcourt Brace Jovanovich, Incorporated.

The Psychological Corporation (1990). *Stanford Achievement Test Series, eighth edition: technical data report*. Atlanta, GA: Harcourt Brace Jovanovich, Incorporated.

The Psychological Corporation (1993). *Metropolitan Achievement Tests seventh edition technical manual spring data.* Atlanta, GA: Harcourt Brace Jovanovich, Incorporated.

Pullias, E. V. (1941). Review of the Metropolitan Achievement Tests. In O. K. Buros' (Ed.) *The nineteen forty mental measurements yearbook*. (pp. 27-29). Highland Park, NJ: The Mental Measurements Yearbook.

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

Reilly, R. R. and Lewis, E. L. (1983). *Educational Psychology*. New York, NY: Macmillan Publishing.

Remmers, H. H., and Gage, N. L. (1955). *Educational measurement and evaluation.* New York, NY: Harper and Brothers.

Remmers, H. H., Gage, N. L., and Rummel, J. F. (1965). *A practical introduction to measurement and evaluation.* New York, NY: Harper and Row.

Resnick, D. P. (1981). Testing in America: a supportive environment. *Phi Delta Kappan*. 62:9. 625-628.

Resnick, L. B. (1981). Introduction: research to inform a debate. *Phi Delta Kappan*. 62:9. 623-624.

Resnick, L. B. (1999). *Reflections on the future of NAEP: instrument for monitoring or for accountability?* Los Angeles, CA: Center for the Study of Evaluation.

Riverside Publishing. (2004). *A distinguished history*. Retrieved August 31, 2005 from http://www.riverpub.com/about/history.html .

Robinson, G. E. and Brandon, D. P. (1994). *NAEP test scores: should they be used to compare and rank state educational quality?* Arlington, VA: Educational Research Service.

Rogers, B. G. (1985). Review of the California Achievement Tests. In J. V. Mitchell, Jr.'s (Ed.) *The ninth mental measurements yearbook.* (pp. 243-246). Lincoln, NE: The University of Nebraska Press.

Ruch, G. M. and Terman, L. M. (1926). *Stanford Achievement Test manual of directions*. Yonkers-on-Hudson, NY: World Book Company.

Schindler, A. W. (1953). Review of the California Achievement Test. In O. K. Buros' (Ed.) *The fourth mental measurements yearbook.* (pp. 6-7). Highland Park, NJ: The Gryphon Press.

Shanker, A. (1990). The social and educational dilemmas of test use. In G. Anrig's (Ed.) *The uses of standardized tests in American education*. (pp. 1-13). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan,* 73:11, 232-238.

Shores, J. H. (1953). Review of the California Achievement Tests. In O. K. Buros' (Ed.) *The fourth mental measurements yearbook*. (pp. 7-9). Highland Park, NJ: The Gryphon Press.

Sokal, M. M. (1987). James McKeen Cattell and mental anthropometry: nineteenth-century science and reform and the origins of psychological testing. In M. M. Sokal's (Ed.) *Psychological testing and American society: 1890-1930*. (pp. 21-45). New Brunswick, NJ: Rutgers University Press.

Spies, R. A. and Plake, B. S., eds. (2005). *The sixteenth mental measurements yearbook.* Lincoln, NE: University of Nebraska Press.

Stoker, H. (1992). Review of the Stanford Achievement Test. In J. C. Impara and B. S. Plake's *The eleventh mental measurements yearbook*. (pp. 863-865). Lincoln, NE: University of Nebraska Press.

Tanner, D. and Tanner, L. (1980). *Curriculum development: theory into practice.* Columbus, OH: Merrill.

Tanner, D. and Tanner, L. (1990). *History of the school curriculum*. New York, NY: Macmillan Publishing Company.

Tanner, Laurel. (1983). Curriculum history and educational leadership. *Educational Leadership*, 41:3, 38-42.

Terman, L. M. (1916). *The measurement of intelligence*. Atlanta, GA: Houghton Mifflin Company.

Terman, L. M. (1919). *The intelligence of school children*. Atlanta, GA: Houghton Mifflin Company.

Terman, L. M. (1923). *Intelligence tests and school reorganization*. Yonkers-on-Hudson, New York, NY: World Book Company.

Thorndike, E. L. , Bregman, E. O., Cobb, M. V., Woodyard, Ella, et al. (1927). *The measurement of intelligence*. New York, NY: Teachers College.

Traxler, A. E. (1944). Individual evaluation. In American Council on Education Studies' *New directions for measurement and guidance*. (pp. 16-34). Washington, DC: American Council on Education.

Tyler, R. W. (1944). Admission and articulation based on study of the individual. In American Council on Education Studies' *New directions for measurement and guidance*. (pp. 1-15). Washington, DC: American Council on Education.

Tyler, R. W. (1971). Accountability in perspective. In L. M. Messinger and R. W. Tyler's (Eds.) *Accountability in education*. (pp. 1-14). Worthington, OH: Charles A. Jones Publishing Company.

United States Department of Education. (2002, March). *Public law 107-110*. Retrieved April 2, 2003, from http://www.ed.gov/legislation/ESEA02/107-110.pdf

University of Iowa College of Education Iowa Testing Programs. (2007). *Purposes of the ITBS Batteries, Levels 5-8*. Retrieved July 5, 2007, from http://www.education.uiowa.edu/itp/itbs/itbs_about_5-8_prp.htm .

Valentine, J. A. (1987). *The College Board and the school curriculum: a history of the College Board's influence on the substance and standards of American education, 1900-1980*. New York, NY: College Entrance Examination Board.

Vinovskis, M. A. (1998). *Overseeing the nation's report card: the creation and evolution of the National Assessment Governing Board*. Washington, DC: United States Department of Education.

Wardrop, J. L. (1976). *Standardized testing in the schools: uses and roles*. Monterey, CA: Brooks/Cole Publishing Company.

Wardrop, J. L. (1989). Review of the California Achievement Tests. In J. C. Conoley and J. J. Kramer's (Eds.) *The tenth mental measurement yearbook*. (pp. 128-133). Lincoln, NE: University of Nebraska Press.

Whipple, G. M. Ed. (1926). *The twenty-sixth yearbook of the National Society for the Study of Education: the foundations and technique of curriculum-construction*. Bloomington, IL: Public School Publishing Company.

Willson, V. L. (1985). Review of the California Achievement Tests. In J. V. Mitchell, Jr.'s (Ed.) *The ninth mental measurements yearbook.* (pp. 246-248). Lincoln, NE: The University of Nebraska Press.

Wolf, R. M. (1978). Review of the Metropolitan Achievement Tests. In O. K. Buros' (Ed.) *The eighth mental measurements yearbook.* (pp. 67-69). Highland Park, NJ: The Gryphon Press.

Womer, F. B. (1978). Review of the California Achievement Tests. In O. K. Buros' (Ed.), *The eighth mental measurements yearbook.* (pp. 37-39). Highland Park, NJ: The Gryphon Press.

Yoakum, C. S. and Yerkes, R. M. eds. (1920). *Army mental tests.* New York, NY: Henry Holt and Company.