The Rise of Standardized Educational Testing in the U.S.: A Bibliographic Overview

Torin Monahan
www.torinmonahan.com
Rensselaer Polytechnic Institute
Department of Science and Technology Studies
December 1998

Abstract

By providing an overview of some of the texts on the history of educational testing in the U.S., this bibliographic essay seeks to convey a sense of the diversity among viewpoints on testing and an appreciation for the historical context and controversies of test development. Overall, the works commented on in this essay each refer back to the unique historical context that allowed testing to flourish in American school systems. The Progressive reform movement, the emergence of psychology as a profession (especially with its ties to the U.S. army), and trust in scientific expertise and quantification each contributed to the widespread incorporation of standardized tests for educational purposes. Educators, policy makers, students, and others are still struggling to sort through the implications and complexities of such testing.

Introduction

The history of standardized educational testing in the United States finds its roots in the interstices of the rise of progressive reform, the emergence of psychology as a profession, the bureaucratization and Taylorization of education and other life-worlds, and the increasing dependency upon scientific expertise and quantification. Many of these forces overlap in their driving rationales, but the context they created provided educators and policy makers with a mechanism for dealing with overwhelming numbers of students in the face of mass immigration and compulsory schooling; this context, which made the problems of lack of accountability and efficiency visible, also pointed the way toward solutions through scientific expertise which could streamline the business of education by determining student intelligence or achievement through

tests and then placing students on appropriate learning tracks in light of those test results. This bibliographic essay will sort through many of the texts that deal with the historical origins and development of, as well as controversies over, standardized educational testing in the United States. Given the complex, heterogeneous nature of this testing phenomenon, this essay does not aim at comprehensiveness but instead intends to offer the reader an appreciation of the diverse ways in which testing still connects to and informs many aspects of our social, economic, and political realities.

History of Intelligence Testing

Stephen Jay Gould's *The Mismeasure of Man* (1981) is perhaps the most popular book on the controversies surrounding the origins of intelligence testing – it won the National Book Critics Circle Award for general nonfiction in 1981. As the book's title indicates, Gould focuses on cases of tests that were misused in order for the testers to achieve data that would support their often racist presuppositions. The results of these tests were then used, in some cases, to the harm of individuals. Gould intends his book to be a caveat for all scientists, and those interested in using the findings of science, to first acknowledge the influence of social norms and values upon scientific inquiry and second to make efforts to overcome the deleterious effects of social influences upon science. Gould writes: "science must be understood as a social phenomenon, a gutsy, human enterprise, not the work of robots programed to collect pure information . . . Science, since people must do it, is a socially embedded activity" (Gould 21).

Some of the science controversies covered by Gould include subjective measurement in nineteenth-century craniology, eugenic influences on hereditarian theories of IQ, and fabricated data on separated identical twins to "prove" that intelligence is transferred genetically.

Throughout these examples, Gould illustrates ways in which what he calls "the allure of numbers" gained vogue with the human sciences and lent culturally biased judgments an aura of scientific objectivity. In fact, the appearance of the objective nature of numbers clouded scientists minds so that they became unaware of subjective influences upon their work: "They believe in their own objectivity, and fail to discern the prejudice that leads them to one interpretation among many consistent with their numbers" (106). While one may read this apology for misguided scientists as overly generous, it does convey a sense of danger associated with numerical interpretations – even experts succumb to the myth of objective numerical interpretation. This theme of the corruptibility of numerical data continues throughout the literature on psychological intelligence testing.

In *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890-1930* (1992), JoAnne Brown analyzes how Progressive Era psychologists, such as Robert Mearns Yerkes, Lewis Madison Terman, Henry Herbert Goddard, and Edward Lee Thorndike, manipulated language to gain hegemony over education evaluation. Not only did these psychologists rely upon the rhetoric of scientific objectivity, but they also drew upon metaphors from engineering and medicine to legitimize the methods of this emerging profession. This linguistic turn allowed psychologists to argue for a superior understanding of student potential than teachers relying on qualitative descriptions based on first-hand knowledge could muster. In other words, teachers' subjectivity interfered with their evaluation of students; psychologists had access to numerical truth to guide their interpretations. Brown relates: "The metaphors of medicine and engineering both established cultural authority of quantitative methods and universal norms over qualitative methods and firsthand, local knowledge" (Brown 9). So, the introduction of psychology-based intelligence tests into schools was not only founded

upon efficiency or need; one might say that in order to legitimize their profession, psychologists capitalized upon an increasing trust in numerical objectivity to cultivate a need for what they professed to offer: objective measurements of student intelligence. Educators fought to retain control over educational policy and pedagogy, but also welcomed psychological techniques as a means for dealing with the practical difficulties of education – mass immigration[1], compulsory schooling[2], low teacher pay, and few trained teachers.

Brown explains that psychologists veered away from overly scientific language because they required public support and public understanding of the basics of their enterprise in order to achieve popular acceptance – their work needed to be practical. At the same time, psychologists needed to assert their unique authority over the area of intelligence testing. Brown argues that psychology gained professionalization, in part, by successfully combining contradictory metaphors of popularity and monopoly (13). They argued for specialized knowledge over their field by using medical metaphors and argued for the numerical efficiency by using engineering metaphors. Through this metaphorical combination, psychologists secured dominance over the field of educational testing: they alone possessed the expertise to manipulate the tests that would reveal popularly accessible numbers. Finally, despite the standardization of intelligence quotient (IQ) tests to a subjective norm, psychologists were able to construct, out of individual tests, the appearance of an objective "universal standard." The popularization of IQ tests served, then, to advertise and to popularize the profession of psychology.

---

[1] "18.2 million immigrants arrived in the United States between 1890 and 1920, compared with only 10 million between 1860 and 1890 . . . From 1899 to 1914 New York City school enrollments jumped 60 percent; in the middle of this period nearly three quarters of New York City schoolchildren had foreign-born fathers. In smaller cities, too, most children came from immigrant families" (Brown 47).

[2] Most states had passed compulsory education laws by 1900 in an attempt to "Americanize" new immigrants (Sokal 12).

Where Brown seeks to explain the metaphorical strategies employed by psychologists, in *Schools As Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890-1930* (1988), Paul Davis Chapman charts the effects of intelligence testing upon school systems and upon students. Both Brown and Chapman focus on the years 1890-1930 because this time period roughly corresponds to the rise and decline of the progressive movement in the United States, along with an increased rate in the number of immigrants. According to Chapman, intelligence testing and the classification of students assisted schools in completing a transformation that was already underway prior to these new technologies and techniques – a transformation from common, decentralized schools to highly structured, differentiated schools. *Schools As Sorters* offers an explanation for the introduction of tests into schools yet also relates ways in which concomitant school restructuring altered social perceptions of students and the goals of education.

Chapman proffers three main reasons for the rapid incorporation of intelligence testing by schools. First, the emergence of a new network of professionals, psychologists and school administrators, joined forces with national organizations[3], philanthropic foundations, and educational publishers to mandate testing for classifying students. Second, as Brown also indicated, schools adopted tests to assist them with administrative and pedagogical crises in the face of increased enrollments, diverse student populations, and compulsory education laws. Finally, intelligence tests resonated with the national values of the Progressive Era: "University professors and school people alike saw the tests as the 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" (Chapman 5).

---

[3] National Education Association, National Society for the Study of Education, National

As a result of widespread test incorporation, schools adopted a new social function of sorting students and tracking their achievement through what appeared to be scientific, rational, and objective means. New categories were created for students: Average pupils, Non-English-Speaking Immigrants, Juvenile Delinquents, Feebleminded, Retarded, Epileptics, Gifted, Cripple, and so on. Schools also restructured to accommodate for the newfound variations in student populations[4], adding fundamental levels in elementary school, an intermediate level between elementary school and high school with some differentiations in courses, junior college, and expanded professional schools (53-54). As advocates of social justice would attack in the uprising against standardized tests in the 1970s, the false aura of scientific objectivity surrounding these new social categories aided the institutionalization of race and class biases; furthermore, the psychological effects of negative student classification upon students often created self-fulfilling prophecies of failure.

Kurt Danziger, in *Constructing the Subject: Historical origins of psychological research* (1990), offers to this field of psychological testing and education a critical look at the social construction of psychological knowledge. Whereas Brown analyzed the metaphorical constructions used by psychologists to legitimate their profession to an outside audience, Danziger delineates ways in which the internal mechanisms of psychological knowledge production both constrained and determined the types of truths psychologists could construct. While psychologists create technologies and techniques for scientific inquiry (tests, serial lists, rating scales, etc.), "[w]hatever guesses are made about the natural world are totally constrained by this world of artifacts" (Danziger 2). Moreover, not only do human values guide the

---

Research Council, and the U.S. Bureau of Education.

[4] Note: Aided by subjective evaluations, this restructuring was already underway prior to the incorporation of intelligence testing by schools.

interpretation of test findings, these values also inform the means of producing that data. Danziger calls for a recognition of the limits of empirical and methodological rationalism for explaining psychological inquiry – there are social as well as logical dimensions of psychological research activities. The most obvious of social dimension of psychological testing is the need for tested individuals to collaborate (or cooperate) with test administers; both a sense of need and formal mechanisms for testing must exist or be created between social actors for the enterprise to succeed.

Rather than concentrate upon individual actors in the development of psychology as a profession, however, Danziger emphasizes synchronic and diachronic patterns of variation in the investigative practices of psychology's practitioners. While Danziger's discussion of patterns of variation allows him to apply with ease the framework of social constructivism, and specifically actor-network theory, to the development of psychology, his reluctance to chart specific actors' contributions to the construction of psychological knowledge diminishes the practical value of this text; his over reliance on theoretical generality distances his text from both his historical topic and his reader.

Franz Samelson's essay

"Was Early Mental Testing:
(a) Racist Inspired,
(b) Objective Science,
(c) A Technology for Democracy,
(d) The Origin of the Multiple-Choice Exams,
(e) None of the Above?
(Mark the RIGHT Answer)" (1987)

tracks multiple-choice testing technology back to its roots. Prior to 1914, this type of testing technology did not exist; by 1921, over 2 million U.S. soldiers and 3 million U.S. school children took multiple-choice tests (116). These tests were developed under the aegis of Robert M.

Yerkes through the American Psychological Association's Committee on Methods of Psychological Examining of Recruits. Arthur S. Otis, however, was the originator of this technology. Working in Yerkes' psychological division, Otis devised the tests and gave them to Terman, who in turn presented them to the APA committee. The tests then became the basis for the Army Alpha test (117).

Samelson continues to explain that while quickly adapted to and adopted by the military, educators retained some skepticism about the effectiveness of these tests for school children. They worried specifically that multiple-choice tests would encourage guesswork and reduce independent thinking – students would only need to focus on fragmented factual information and would not develop any sophisticated understanding of the material. In spite of these concerns, educators were pressured to embrace these "modern" scientific practices. By the late 1920s, school boards and test publishers had joined forces, and those educators critical of these tests could not withstand the ideological and bureaucratic forces pushing for test adoption. In 1926, for instance, the College Entrance Examination Board added Scholastic Aptitude Test (SAT) scores as a criteria for student selection (122).

John Carson's essay "Army Alpha, Army Brass, and the Search for Army Intelligence" (1993) traces the ways that the emerging profession of psychology tailored its intelligence tests to the needs of the U.S. army in World War One. Prior to 1917, the core group of army officers numbered around 6,000, and most of those were trained at West Point – military culture was relatively small and intimate. By 1918, due to mobilization for the war, the army had over 200,000 officers and faced a crisis in determining which new recruits were officer material (Carson 282). Psychologists, most notably Robert Yerkes, stepped in with a plan to administer a ten-minute intelligence test to new recruits exhibiting unsatisfactory behavior in order to assist

the army in determining which men were unfit to serve. The military accepted this plan on several conditions: that tests determine excellence as well as deficiency (to aid them in identifying new officers) and that tests be administered to all new recruits (to assist the military in classifying soldiers). This massive testing required Yerkes to devise methods for quick, objective administering and scoring of tests and to devise tests that would foil cheating. This gave rise to the first army intelligence scale: Army α.

Since accommodations made by psychologists to the needs of the military helped shape the character of this new profession, and specifically the sub-field of psychometrics, one can reasonably conclude that school tests devised by psychologists after WWI also carried traces of this military flavor. The greatest of psychologists' allowances to the military were the elimination of middlemen to interpret the findings: "professional judgment was subordinated to objective determination and statistical manipulation" (287). So while the army used Army α findings to balance out their companies, they also opted to ignore tests that found men unfit to serve and to retain their own criteria for overriding test findings: officer opinions. In school settings, this lack of interpretation would lend the appearance of even greater validity to the numerical findings of intelligence tests – "the numbers speak for themselves." Besides an increased reliance on unmediated numbers, psychologists also adapted their testing procedures to take on a militaristic appearance, right down to yelling commands at test takers (299) and incorporating military language into the tests (300).

Theodore M. Porter echoes this theme of the colonization of life-worlds through quantification in *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (1995). Porter relates how rules and methods function to replace selfhood or individual volition and lead to the objectification or dehumanization of the people described by numbers (Porter 76-77).

Porter also notes the uniquely American nature of standardized tests; they grew out of military and industrial environments with promises of mechanical objectivity and found purchase in the soil of public schools (209). Educators used standardized tests primarily for sorting students into differentiated, educational tracks – such as academic, vocational, or commercial.

Michel Foucault's writings on the power of technological systems for depriving individuals of liberty offer another critique of the effects of school systems and their testing apparatuses upon children. In *Discipline and Punish: The Birth of the Prison* (1977), Foucault calls attention to the structural similarities in technological systems, such as schools, prisons, and hospitals, and comments upon the combined attenuating effects of these institutions upon individual autonomy and liberty. The dehumanizing effects of quantification articulated by Porter take on a more sinister and insidious form in Foucault's text:

> [The prison] had already been constituted outside the legal apparatus when, throughout the social body, procedures were being elaborated for distributing individuals, fixing them in space, classifying them, extracting from them the maximum in time and forces, training their bodies, coding their continuous behavior, maintaining them in perfect visibility, forming around them an apparatus of observation, registration and recording, constituting on them a body of knowledge that is accumulated and centralized. (Foucault 231)

Foucault argues that people learn to discipline and regulate their own behavior in response to the demands of conformity placed upon them by technological systems – people embody the prison. School systems, with their regimented tasks, continuous observation, and use of testing as a means of accountability and classification, contribute to this seldom recognized deprivation of autonomy and liberty through embodied regulation.

General Testing Controversies

Paul Davis Chapman locates the temporal origin of widespread debates about school testing near the end of WWI: "When testing and tracking first became widespread in the twenties, psychologists, educators, and journalists engaged in a furious debate about intelligence tests" (Chapman 2). In 1962, Banesh Hoffman's provided a minor flair up in the school testing controversy with *The Tyranny of Testing* (1962), but nothing like that following Arthur R. Jensen's *Harvard Educational Review* article "How Much Can We Boost IQ and Scholastic Achievement" (1969). This article was then re-printed in a follow-up, full-length book by Jensen entitled *Genetics and Education* (1972).

Jensen's basic argument is that "[c]ompensatory education has been tried and it apparently failed" (Jensen 69). Jensen's tone is academic and guarded, but his message is clear: programs designed to assist minority groups and the economically disadvantaged in bringing scholastic performance up to par have failed because a physiological difference exists between races – whites are more intelligent. In the political context of civil rights movements, this eugenic-like message was met with open hostility, including threats upon Jensen's life.

The discovery of a scandal with one of the fathers of hereditary-based intelligence, Sir Cyril Burt, added additional fuel to this conflagration. As covered by Gould in *The Mismeasure of Man*, Burt fabricated data on identical twin research to "prove" the hereditarian theory of intelligence transfer (he conducted his research from the early part of the twentieth-century up until 1966). The discovery of this fraudulent research in 1973 had special implications for the Jensen debate because "Arthur Jensen used Sir Cyril's figures as the most important datum in his notorious article . . ." (Gould 265). Leon Kamin, the Princeton researcher and psychology

professor who exposed Burt's fabrications, felt appalled that this racist work contributed to unjust social policy by influencing the U.S. Immigration Restriction Act of 1924 (Fancher 208), and he began to publish works on the racist proclivities of early intelligence testers; Kamin's most influential publication was *The Science and Politics of I.Q.* (1974). Despite the fact that these early psychologists testified before Congress arguing for immigration restriction, other researchers have found no definitive evidence that these psychologists influenced the immigration law that followed (Sokal 7).

Raymond E. Fancher, in *The Intelligence Men: Makers of the IQ Controversy* (1985), aspires to quell these polemically charged IQ controversies and the wrongful persecution of some scientists by simplifying the terms of these controversies into two distinct but interrelated issues: the nature/nurture issue (not one or the other, but to what degree is IQ shaped by each) and the definition of intelligence (tests measure something, but do not measure pure intellectual ability). Fancher invests a great deal of faith into the practical and logical outcomes of his simplified (and simplistic) analysis: "This book attempts to ease if not remove the bewilderment [of conflicting interpretations about IQ testing] by examining the two aspects of the IQ controversy from a biographical and historical perspective" (xiv). This book, with its trust that the reasonableness of people will provide a means for overcoming dispute and its implied belief in the eventual perfectability of tests, offers a conservative foil to the many liberal texts attacking tests as mechanisms for institutionalizing discrimination and ensuring corporate hegemony over knowledge production.

Throughout his biographical accounts of the fathers of psychological testing, Fancher adopts the viewpoint that while the tests may not have been perfect, with few exceptions, the science behind them was sound. So, any misuse of these technologies for discrimination must be

attributed to social prejudice, not to scientific error. As with his division of the reasons for confusion over testing, Fancher continues with his penchant for separation and classification to shield science from blame for the misuse of its neutral technologies. In an nimble rhetorical move, Fancher acknowledges cases of flawed scientific analyses, but protects his vision of the purity of science by suggesting that since these cases were flawed, they were not real science (Fancher 239). The final passage of this book neatly demarcates the arena where knowledge issues should be debated; it also betrays Fancher's elitism: "More powerful than any available scientific data, these different experiences [of the individual formative effects of nature or nurture] will continue to produce a wide variety of views – both in the scientists who will continue actually to debate the IQ controversy, and in the public who will continue to observe it with interest" (241).

In his essay "History of Educational Testing" (1982), Daniel Resnick offers a historical overview of standardized testing in American educational settings. He tracks these tests, which allow for comparisons between student populations, from their first use in 1840 up until after World War Two. Resnick stresses that the widespread use of these tests as a means of evaluating student capabilities and achievement is unique to the U.S.. This country's ideology of individuality and faith in measurement devices, combined with school system crises at the turn of the century, created an environment conducive to standardized testing. In regard to current controversies surrounding testing, Resnick avers that the three major agents of the testing enterprise – "applied psychology, school administration, and the publishing industry" (Resnick 191) – have too firm a hold over education evaluation to let go. Moreover, whenever a crisis is identified within schools, the public usually calls for stronger means of accountability through testing. Resnick concludes: "The present waves of controversy would have to wash very high to

erode a base of use and support that has grown considerably in size and character over the past three-quarters of a century" (174).

H. J. Butcher's *Human Intelligence: Its Nature and Assessment* (1970) does not devote much space to issues of educational testing, but it does provide insight into traditional psychologists' perspectives on the merits and proper role of educational testing. Focusing mainly upon theories of the development of intelligence and types of intellectual abilities, Butcher indicates that psychology should veer away from the detailed studies of individual differences in intelligence, such as the fields of psychometrics and experimental psychology, and instead concentrate upon "the basic laws of cognitive functioning" (Butcher 10). If the results of tests are only one indicator of intelligence or achievement, which Butcher avers, then schools should place greater emphasis upon cumulative grade point averages than upon individual tests for purposes of student assessment (282). Furthermore, Butcher concludes that the predictive merits of tests for determining student success are uncertain (282); critics of the Educational Testing Service (ETS) would later harp upon this very issue of non-predictability.

Erness Bright Brody and Nathan Brody extend this commentary on the merits of intelligence testing for predicting student progress in *Intelligence: Nature, Determinants, and Consequences* (1976). The authors claim that "data exist that suggest success in school acts as a determinant of intelligence . . ." (Brody & Brody 88) rather than a consequence of intelligence. Although the authors focus more upon arguing that school achievement and intelligence test scores do measure separate things that can be compared, they also acknowledge the messy social aspects of both measures. That is, success in school and intelligence scores may have more to do with parents' education and socioeconomic status than innate ability (90). Moreover, the high rankings on a 1945 table of IQ scores sorted by profession correspond conspicuously with

prestigious, well paying occupations; whereas lower rankings correspond with less prestigious, lower paying occupations (93-94). The authors intended to draw a correlation between IQ scores and social prestige, but the table also invites readings that IQ tests measure social prestige rather than intelligence. For example, accountants, lawyers, and engineers are highest on the list, salesmen, musicians, and artists are in the middle range, and teamsters, farmers, and miners are at the bottom.

Brody and Brody conclude with some salient comments on the use of intelligence tests in schools: "Although it is difficult to ascertain the positive value of intelligence tests as used in the public schools, it is not difficult to suggest possible negative consequences that follow from the use of tests" (210). Given IQ tests' openness for misuse and their inability to provide educators with information that could not be gained through other means, the authors recommend the discontinuation of IQ tests in public schools. The tests, in fact, have little relevance to schools because educators already know, first-hand, about the achievement of their pupils (209). This conclusion, by the authors, betrays an assumption of adequate educational resources for schools, but even if that is not the case, intelligence testing would not be a panacea for school financial difficulties.

ETS Controversies

In the 1970s, an increased skepticism about the efficacy and usefulness of student achievement and intelligence tests for schools brought about numerous publications protesting testing and perceived testing monopolies. These publications both mirrored real social unrest over these issues, such as political protests, and helped catalyze some important legislative

changes concerning the testing industry – most notably, the Truth-in-Testing law[5] passed by New York in 1979. The most influential of these publications leading up to the Truth-in-Testing law were *Education and the Cult of Efficiency* (1962) by Raymond Callahan, *The Tyranny of Testing* (1962) by Banesh Hoffman, *Schools in an Age of Mass Culture* (1965) by Willis Rudy, *Edward L. Thorndike: The Sane Positivist* (1968) by Geraldine Joncich, "Education and the Corporate Order" (1972) by David K. Cohen and Marvin Lazerson, *The One Best System* (1974) by David Tyack, *Schooling in Capitalist America* (1976) by Samuel Bowles and Herbert Gintis Cronbach, and "Testing for Order and Control in the Liberal Corporate State" (1976) by Clarence Karier. The biggest opponent of Truth-in-Testing legislation was Princeton's Educational Testing Service (ETS).

ETS clearly states their testing rationale in *ETS Builds a Test* (1959). In this document, which is more like an information pamphlet than a book, ETS explains that they devise tests with the aim of capturing existing student talents; they also engage in testing research and provide advisory services. Whereas ETS opponents claim that this company creates a need where none exists, ETS asserts the opposite: "Tests are built to meet the needs of education" (ETS 3). Using rhetoric that could have come straight out of Porter's *Trust in Numbers*, ETS claims that their tests are completely objective and rational; they also debunk any arguments that tests do not really measure achievement and intelligence: "It is erroneously believed in some quarters that objective items, although satisfactory for assessing knowledge of specific facts and information, cannot get at the quality of the test-taker's thinking and judgment . . ." (11). One example they offer to support this claim is an aural test question (for foreign language students) in the form of

[5] "[R]equires higher education admission testing companies to disclose internal studies of their tests, give candidates factual information about what test scores mean, and disclose test questions and correct answers after scores have been reported" (Nairn 136).

a picture. This drawing depicts a white woman, dressed in an apron, carrying a cake to a man, wearing a tie, sitting at a dinning room table. Both the woman and the man are smiling. The woman does not notice a cat that is directly under her upraised foot. ETS asks students to describe "what the picture tells" (13). Evidently, ETS has not distilled Western culture out of their objective tests for foreign students. One final claim from this document will raise the issue of scientific expertise called into question by ETS critics: ETS asserts that one of their greatest tasks is conveying test findings to educators in a simplified form without sacrificing accuracy (22).

Several texts published after the passing of the Truth-in-Testing law relate the details of the controversy and continue a critique of standardized testing and testing organizations such as ETS. The most notable of these texts are *Bias in Mental Testing* (1980) by Arthur Jensen, *The Reign of ETS: The corporation that makes up minds* (1980) by Allan Nairn, *The Testing Trap* (1981) by Andrew J. Strenio, and *None of the Above: Behind the Myth of Scholastic Aptitude* (1985) by David Owen. Since *The Reign of ETS* is the most comprehensive of these texts, I will devote some time to discussing it here.

The Reign of ETS, written by Allan Nairn, was published as a "Ralph Nader Report on the Educational Testing Service." In Nader's preface to this report, he underlines the key reasons for undertaking such a study. First, students often interpret low test scores as revealed *truths* about their insufficiencies; so, low scores can become self-fulfilling prophesies for these students. Parents, teachers, and peers also reify the importance of these scores and treat them as accurate indicators of student ability. Second, scientific expertise should not go unquestioned, especially not when it has such a life-shaping impact on so many people. Nader avows:

When psychometricians, as with other specialists, allege complexity and assert that test

> consumers are not really qualified to question the testing sovereignty that affects their lives so profoundly, it is time for them to evaluate their testers" (x).

Finally, broader approaches to student evaluation are needed – ones that utilize multiple-criteria and take into consideration student plurality. Having a single gate-keeper allots too much power to one organization.

The Reign of ETS traces the origins of ETS back to military intelligence testing of the kind described by Carson. Henry Chauncey, an examiner for the U.S. Navy and a former dean of Harvard, joined forces with Devereux Colt Josephs, the president of New York Life Insurance Company and a former president of the Carnegie Corporation, to charter the ETS corporation in 1947. The ETS mission statement openly betrays a militaristic Cold War ideology: "To serve governmental agencies by providing tests and related services in their educational and training efforts and particularly in time of national emergency to serve the federal government in other activities . . ." (Nairn 2-3). Even further, the Bylaws of ETS promise that it will continue to function even in the wake of an atomic holocaust (294)! The corporation became enormously successful. By 1972, it rivaled the CIA in information collection, storage, and retrieval (28-9) and had more customers than Ford and General Motors combined (29). Finally, for test administering purposes, ETS functions like a well-oiled military machine, complete with synchronized watches, emergency phone services, security provided by ex-CIA and -FBI agents, fingerprinting, etc. (29-31).

Nairn also draws attention to ETS's lack of accountability combined with their enormous, tax-exempt annual income ($94 million in 1980). ETS retains non-profit status because it funnels its income back into the corporation, has no stock holders, and uses other tax tricks (40). Their non-profit and non-academic status also removed them, up until the Truth-in-Testing law,

from any sort of accountability. Furthermore, ETS's intimate relationship with the College Board (or College Entrance Examination Board) has assisted them in creating a demand for their products (tests) that students bear the financial burden of, not universities. Nairn characterizes this state of affairs as "Student consumers in captivity" (260).

In the 1970s, student organizations began lobbying for more rights in this testing market. Organizations such as New York Student Public Interest Research Group (NYPIRG) demanded what they perceived to be their rights as consumers: accurate scoring, no testing errors, access to their tests, etc.. They also questioned both ETS's objectivity and monopoly over the testing market. If test-prep classes could improve scores, then ETS tests really were not objective tests of student ability; those with the money to take such courses and to re-take tests would fair better than those who could not. Nairn describes this as an institutionalized class and race bias that grew out of the discriminatory testing used in to keep Mediterranean, Slavic, and Irish immigrants out of the U.S. in the 1920s (162-3). Once ETS was forced to open up their data vaults, it was apparent why they had argued for secrecy and hid behind a veil of scientific expertise. Internal ETS studies revealed their awareness of anxiety influencing test scores and minority biases embedded in tests (87, 113-115, 129).

Psychological Testing and American Society: 1890-1930 (1987), edited by Michael M. Sokal, provides a stimulating collection of historical essays designed to illuminate the present, and its many controversies, with new readings of the past (5). The introduction by Sokal first lays the groundwork for the current contentious climate by briefly explaining the viewpoints of individual researchers and emphasizing the "publicness" of the nature/nurture testing debate – even Dan Rather ran a CBS documentary of "The I.Q. Myth" in 1975. As in *The Reign of ETS*, Sokal recounts the conditions of the current "ETS state" where parents, teachers, administrators,

and students each buy into the testing system without ever questioning its merits. Sokal delimits the criticisms of the ETS state into three categories: testers are too close to their enterprise to filter out biases (they may even be self-consciously self-serving), tests reinforce class boundaries (a Marxist critique), and tests may be irrelevant or insufficient for predicting performance (Sokal 4). To be fair, Sokal also recounts ETS's counter arguments: if tests are imperfect, more research needs to be done; tests create opportunities for students; tests may not be perfect, but they help more students than they hurt (5).

Evaluation Policy

Educational Testing and Evaluating: Design, Analysis, and Policy (1980), edited by Eva L. Baker and Edys S. Quellmalz, offers a collection of essays on what they identify as an under-investigated topic: educational policy. Baker cites the plurality of American educational systems and student populations as the main reason for this deficiency. The essays in this volume, which grew out of a Measurement and Methodology in Education conference hosted by UCLA's Center for the Study of Evaluation (CSE) in 1978, aspire to improve educational policy through a critical inquiry into evaluation practices. The authors crafted the essays with readability in mind, so they kept the presentation of quantitative data to a minimum. One of the key suggestions made by these authors is to gear testing to the instruction that goes on in unique educational contexts rather than basing tests on a stable and general model of student achievement (13). The authors acknowledge the impracticality of this model but also raise what they call the philosophical issues (or problems) of abdicating total responsibility to measurement experts. Baker suggests four criteria for judging test quality: "(1) that a test be meaningful; (2) that its purpose and structure be public; (3) that it gather information as economically as possible; and

(4) that it address skills sensitive or amenable to instruction" (21). In spite of test difficulties and controversies, having some form of accountability is better than none – as long as it is "fair" (30).

David W. Barnett identifies another side of evaluation policy in *Nondiscriminatory Multifactored Assessment: A Sourcebook* (1983). Barnett seeks to elucidate ways of handling the complicated assessment policy and practice procedures for special student populations: those with learning difficulties or psychical impediments to learning. Barnett suggests that a common costly error of dealing with these populations is over-testing. Multifactored assessments[6] can help parents and educators avoid this error and focus more on the children's needs rather than their placement into categories of disability. Recent legal mandates for Individualized Educational Programs and non-traditional methods of evaluation (ones that don't aim to compare student progress), create new challenges for educators that this text addresses by functioning as a reference book on the topic.

In *Educational and Psychological Testing: A Study of the Industry and Its Practitioners* (1972), Milton G. Holmen and Richard Docter review a number of criticisms of standardized testing for educational and vocational uses and then offer some suggestions for ensuring better use of tests. In several passages, the tone adopted by these authors reveals their underlying belief in the potential for the development of neutral and fair tests, provided human misuse or ignorance can be overcome. For example, "Too often, however, the individual assessment of children has been left to partially trained psychologists or teachers who received little or no supervision as they learned to administer tests. It is not surprising, therefore, that we continue to

[6] Multifactored assessments break student performance down into domains, such as language and communication skills or visual-motor and gross-motor skills, that assist educators in structuring individualized educational programs for these students (Barnett 11)..

see examples of poor professional practice" (Holmen and Docter 5). Given this heralding of expertise in the book's introduction, it is not surprising that the suggestions that the authors recommend as being the most important all revolve around better trained test administers, test reviewers, test assessors, and even test publishers (161-164).

Conclusion

By providing an overview of some of the texts on the history of educational testing in the U.S., this bibliographic essay has sought to convey a sense of the diversity among viewpoints on testing and an appreciation for the historical context of test development and controversy. My commentary on texts about the history of intelligence testing, the controversies surrounding educational testing, and the policy challenges raised by testing has not aimed for comprehensiveness – I elide coverage of recent national and international comparative standardized testing, for instance, as well as recent movements by many states to phase-in mandatory standardized tests in conjunction with curriculum reform. I also neglected to survey the literature on vocational testing. Still, this essay does offer an introduction to some of the most influential texts on the history of and controversies over educational testing in the U.S..

While ETS appears to have retained hegemony over the testing market, in spite of Truth-in-Testing legislation, some anti-testing organizations continue to question the need for standardized testing as a means for assessing student achievement[7]. Since my section on testing policy issues deals more with contemporary trends than historical developments, I limited that section to a few works that point to the kinds of policy and praxis issues that testing mandates raise. Overall, the works commented on in this essay each refer back to the unique historical

[7] See www.fairtest.org for updates on current anti-testing movements.

context that allowed testing to flourish in American school systems. The Progressive reform movement, the emergence of psychology as a profession (especially with its ties to the U.S. army), and trust in scientific expertise and quantification each contributed to the widespread incorporation of standardized tests for educational purposes. Educators, policy makers, students, and others are still struggling to sort through the implications and complexities of such testing.

Appendix – Timeline:

Important Events in the Development of Standardized Educational Tests in the U.S.

1840s –	First localized trial of standardized tests.
1882 –	Sir Francis Galton opens to the public an anthropometric measuring lab in U.K.
1890 –	James McKeen Cattell, working under Wilhelm Wundt, coins the term "mental tests."
1904-1911 –	French psychologist Alfred Binet devises new methods for measuring intelligence.
1914 –	Edmund Huey publishes Stanford-Binet scale.
1916 –	Lewis M. Terman produces Stanford Revision of the Binet-Simon Intelligence Scale.
1917-1919 –	Yerkes & Terman developed intelligence tests to assist Army in classifying 1.7 million recruits: Army α.
1919 –	Under National Academy of Sciences, Terman transformed army tests into National Intelligence Tests for schoolchildren.
1924 –	U.S. Immigration Restriction Act.
1925 –	Over 75 tests of general mental ability developed by psychologists.
1947 –	ETS incorporated.
1969 –	Arthur R. Jensen publishes *Harvard Educational Review* article "How Much Can We Boost IQ and Scholastic Achievement."
1974 –	Leon Kamin publishes *The Science and Politics of I.Q.*, debunking Sir Cyril Burts research on the heriditarian theory of intelligence transfer.
Early 1970s –	Judge J. Skelly Wright abolished use of aptitude testing for tracking students because it deprived poor of equal educational opportunities.
1970s –	National Education Association calls for a moratorium on all standardized testing.
1979 –	Truth-in-Testing law passed by New York.

Selected Bibliography

Baker, Eva L. and Edys S. Quellmalz, eds.. *Educational Testing and Evaluation: Design, Analysis, and Policy*. Beverly Hills, California: Sage Publications, 1980.

Barnett, David W. *Nondiscriminatory Multifactored Assessment: A Sourcebook.* New York: Human Sciences Press, Inc., 1983.

Bowles, Samuel and Herbert Gintis. *Schooling in Capitalist America*. New York: Basic Books, 1976.

Brody, Erness Bright and Nathan Brody. *Intelligence: Nature, Determinants, and Consequences*. Educational Psychology. Ed. Allen J. Edwards. New York: Academic Press, 1976.

Brown, JoAnne. *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890_1930.* Princeton, New Jersey: Princeton University Press, 1992.

Butcher, H.J. *Human Intelligence: Its Nature and Assessment*. London: Methuen & Co LTD, 1968.

Callahan, Raymond. *Education and the Cult of Efficiency*. Chicago: University of Chicago Press, 1962.

Carson, John. "Army Alpha, Army Brass, and the Search for Army Intelligence." *Isis* 84 (1993): 278-309.

Chapman, Paul Davis. *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890-1930.* New York: NewYork University Press, 1988.

Cremin, Lawrence. *The Transformation of the School: Progressivism in American Education, 1876-1957*. New York: Knopf, 1961.

Danziger, Kurt. *Constructing the subject: Historical origins of psychological research.* Cambridge Studies in the History of Psychology. Eds. Woodward, William R. and Mitchell G. Ash. New York: Cambridge University Press, 1990.

Educational Testing Service. *ETS Builds a Test*. 1959. Princeton, New Jersey: Educational Testing Service, 1965.

Fancher, Raymond E. *The Intelligence Men, Makers of the IQ Controversy*. New York: W.W. Norton & Company, 1987.

Foucault, Michel. *Discipline and Punish: The Birth of the Prison*. 1977. 2nd Vintage Books edition. New York: Vintage Books, 1995.

Gould, Stephen Jay. *The Mismeasure of Man*. 1981. Rev. and expanded ed. New York: W.W. Norton & Company, 1996.

Hoffman, Banesh. *The Tyranny of Testing*. New York: Crowell-Collier Press, 1962.

Holmen, Milton G. and Richard Docter. *Educational and Psychological Testing: A Study of the Industry and Its Practices*. New York: Russell Sage Foundation, 1972.

Jensen, Arthur R. *Genetics and Education*. New York: Harper & Row, Publishers, 1972.

---. *Bias in Mental Testing*. New York: Free Press, 1980.

Joncich, Geraldine. *Edward L. Thorndike: The Sane Positivist*. Middletown, Conn.: Wesleyan University Press, 1968.

Kamin, Leon J. *The Science and Politics of I.Q.* Potomac, Md.: Lawrence Erlbaum Associates, 1974.

Nairn, Allan and Associates. *The Reign of ETS: The corporation that makes up minds*. The Ralph Nader Report on the Educational Testing Service. 1980.

Owen, David. *None of the Above*. Boston: Houghton Mifflin, 1985.

Porter, Theodore M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, New Jersey: Princeton University Press, 1995.

Resnick, Daniel. "History of Educational Testing." *Ability Testing: Uses, Consequences, and Controversies*. Part II: Documentation Section. Alexandra K. Wigdor and Wendell R. Garner, eds. Washington, D.C.: National Academy Press, 1982.

Rudy, Willis. *Schools in an Age of Mass Culture*. Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

Sokal Michael M., ed. *Psychological Testing and American Society: 1890-1930*. London: Rutgers University Press, 1987.

Strenio, Andrew J. *The Testing Trap*. New York: Rawson, Wade Publishers, 1981.

Tyack, David B. *The One Best System: A History of American Urban Education*. Cambridge, Mass.: Harvard University Press, 1974.