Testing and Social Stratification in American Education

Eric Grodsky,[1] John Robert Warren,[2] and Erika Felts[1]

[1]Department of Sociology, University of California, Davis, California 95616; email: esgrodsky@ucdavis.edu, edfelts@ucdavis.edu

[2]Department of Sociology, University of Minnesota, Minneapolis, Minnesota 55455; email: warre046@umn.edu

Annu. Rev. Sociol. 2008. 34:385–404

First published online as a Review in Advance on April 14, 2008

The *Annual Review of Sociology* is online at soc.annualreviews.org

This article's doi: 10.1146/annurev.soc.34.040507.134711

0360-0572/08/0811-0385$20.00

Key Words

standardized tests, socioeconomic status, opportunity to learn

Abstract

We focus on how standardized testing in American education has reflected, reproduced, and transformed social inequalities. We begin by describing inequalities in test score distributions by race/ethnicity, social origins, and gender over time. We then define learning, cognitive ability, and opportunity to learn, each of which influences the results of standardized tests. Next, we offer a brief history of standardized testing's role in American education. We then discuss the relationship between social stratification and measurement issues that arise in the context of standardized testing and the contemporary uses and misuses of standardized testing for diagnostic purposes, accountability, and gatekeeping. We conclude by reflecting on the past, present, and future role of testing in social stratification.

INTRODUCTION

The relationship between standardized testing and stratification in American education has a lengthy and contradictory history. On the one hand, proponents assert that some types of standardized tests undermine social stratification by reflecting the intellectual or academic potential of students rather than their actual achievements, arguing that achievement is shaped by circumstance, whereas potential is not. On the other hand, critics contend that standardized tests are fundamentally flawed, biased against historically disadvantaged groups, and thus have the pernicious effect of reinforcing social advantage and disadvantage.

In this review, we focus on how standardized testing in American education has reflected, reproduced, and transformed social inequalities by race/ethnicity, social origins, and gender. Many readers may equate standardized tests with multiple-choice tests. Standardization (for our purposes) refers not solely to the content of tests or to how tests are scaled but to the way they are administered. Standardized tests present students with identical tasks in environments that are (to the extent possible) uniform, with the hope that variation in test performance is unrelated to variation across test administrations and test administrators. Standardized tests are widely used for a variety of administrative, evaluative, and bureaucratic purposes. In this review, however, we mostly confine our attention to tests of cognitive skills or academic achievement given in the course of the primary, middle, and secondary school years.

Sociological interest in standardized tests and the use of testing as a means of limiting access to social resources is hardly new. In fact, both Sorokin (1959) and Weber (1978) highlight the long-standing use of exams in China to ration access to high-level bureaucratic positions in the public service. Sorokin writes about examinations (and assessment more generally) as a critical mechanism by which modern schools accomplish their work of sorting students into their prospective social positions. Although Sorokin acknowledges the role of schools in actually educating students, he asserts that, from a functionalist standpoint, "*the school is primarily a testing, selecting, and distributing agency*" (Sorokin 1959, p. 188, emphasis in original).

Weber, on the other hand, traces the functional necessity of examinations not to a sorting function demanded of all stratified societies, but instead to the functional demands of the modern bureaucracy. As the bureaucratic organizational form assumed dominance, Weber writes, the "patent of education" came to replace ancestry as a means of legitimate qualification for state office. Education in general, and exams in particular, serve to "limit the supply of candidates for [socially and economically advantageous] positions and to monopolize them for the holders of educational patents" (Weber 1978).

We begin our review by describing inequalities in test score distributions by race/ethnicity, social origins, and gender over time. These empirical patterns drive much of the controversy around testing, at least as it pertains to social stratification. We then define learning, cognitive ability, and opportunity to learn (OTL), each of which influences the results of standardized tests and patterns of social stratification. Next, we offer a brief history of the role of standardized testing in American education (with college entrance examinations serving as an important example). We then discuss the relationship between social stratification and measurement issues that arise in the context of standardized testing and the contemporary uses and misuses of standardized testing for diagnostic purposes, accountability, and gatekeeping. We conclude with some reflections on the past, present, and future role of testing in social stratification.

INEQUALITIES IN STANDARDIZED TEST SCORES

Sociologists would probably not be nearly as interested in standardized test scores were it not for the fact that average test scores vary along familiar lines of social inequality.

Although variation within groups usually trumps variation between groups, differences in mean test scores reflect important differences in home and school resources as well as possible flaws in the tests themselves. In this section, we give an overview of how test scores vary by race/ethnicity, social origins, and gender over time.

Racial/ethnic differences in mean standardized test scores are evident from the earliest years of formal schooling, with African American and Hispanic children scoring below non-Hispanic white children, and with Asian American children scoring above non-Hispanic white children on standardized reading and mathematics tests. These patterns are observed in kindergarten (Fryer & Levitt 2005), elementary school (KewalRamani et al. 2007), and middle and high school (Hedges & Nowell 1999). Similar patterns are observed in results of standardized tests used for college (such as the ACT and SAT) and graduate school admissions (such as the GRE, the GMAT, the LSAT, and the MCAT) (Schmidt & Camara 2004). Racial/ethnic differences in mathematics and reading test scores at the secondary school level declined through the 1980s and stabilized thereafter (Jencks & Phillips 1998, Neal 2006). Black-white differences in SAT scores remained stable between the late 1980s and 2006, whereas Latino-white differences may have increased slightly in math and declined slightly in reading (Kobrin et al. 2007).

Standardized test scores consistently vary by socioeconomic status (SES) as reflected by parental education, family income, or other measures in kindergarten (Fryer & Levitt 2004), elementary school (Rathbun et al. 2004), and middle and high school. SES differences extend to scores on tests used for college and graduate admissions (Camara & Schmidt 1999). Long-term trend data from the National Assessment of Educational Progress (NAEP) suggest that differences in reading and math achievement across levels of parental education were unchanged between 1980 and 2004, with children of college-educated parents outscoring those of high school dropouts by 30 points at age 13 in both subjects, by 30 points in math at age 17, and by almost 40 points in reading at age 17 (Perie et al. 2005).

Given means differences in SES across racial/ethnic groups, part of racial/ethnic group differences in standardized test scores is likely due to socioeconomic differences across groups. However, racial/ethnic variation in standardized test scores persists even among young people from similar socioeconomic backgrounds (Jencks & Phillips 1998). Some evidence suggests that test score gaps actually increase with levels of SES (Gosa & Alexander 2007, Grodsky et al. 2007, Lubienski 2002).

Gender differences in standardized test scores are relatively small, at least compared with racial/ethnic and socioeconomic differences (see Buchmann et al. 2008, in this volume, for a more detailed review of sex difference in academic achievement). In reading, there are small differences between boys and girls in kindergarten and in the early grades (Fryer & Levitt 2005); these modest differences favor girls. In math, small but statistically significant differences favoring boys are evident as early as first grade (Hedges & Nowell 1995). The modest gender differences in mathematics achievement measured on the long-term trend NAEP data have declined over time. In reading, the gender gap continues to favors girls, growing from 5 points among 9-year-olds to 14 points among 17-year-olds (Perie et al. 2005).

Patterns of variation in standardized test scores across groups set the stage for an analysis of the role of standardized tests in reflecting, reproducing, and transforming broader patterns of social inequality. The determinants of the differences we describe above are subject to substantial debate. Some argue that standardized test scores accurately capture variation in the knowledge and skills that different groups of students possess, or even innate differences in their cognitive abilities (Herrnstein & Murray 1994). Others contend that test score variation largely reflects variation in social resources and OTL (Fischer et al. 1996). Still others dispute the evidence itself, arguing that

standardized tests inaccurately characterize differences in underlying skills among groups. To resolve these controversies one must consider the complex relationship between learning, cognitive ability, and OTL, as well as some of the technical underpinnings of test development and scoring.

LEARNING, COGNITIVE ABILITY, AND OPPORTUNITY TO LEARN

Learning is a process of change in which students accumulate knowledge and skills over time. In her review of the sociological literature 40 years ago, Boocock (1966) noted that surprisingly little attention had been paid to learning up to that time. Since then, the empirical literature on learning has evolved substantially, but with relatively few exceptions sociological theories of learning have not. One of those exceptions is the work of Sorensen & Hallinan (1977). They propose a model in which the rate of learning is a nonlinear function of student ability, effort, and OTL. As the authors demonstrate, this model is difficult to identify over long periods because effort, ability, and OTL are to some degree endogenous to one another. In contrast to Sorensen & Hallinan, Kilgore & Pendleton (1993) assert that the relationship among these three factors is additive rather than multiplicative and that effort is endogenous to OTL.

Providing an authoritative definition of cognitive ability or intelligence is challenging. In two symposia held in 1921 and 1986, leading scholars of intelligence offered over two dozen competing definitions (Sternberg et al. 2005). This disagreement has not deterred psychologists and other social scientists from studying intelligence; in Jensen's (1969) words, "[t]here is no point in arguing the question to which there is no answer, the question of what intelligence *really* is." It has, however, allowed the measurement of the concept to drive the concept itself. Although the starting point for many operational definitions of intelligence is *g*, the fundamental function that "all branches of intellectual activity have in common" (Spearman 1904), contemporary intelligence researchers argue for 2, 3, or even as many as 60 hierarchically nested cognitive abilities following from *g* (Carroll 1997).

Intelligence is closely related to problem solving ability and to the rate at which people learn, and it has been conceived of by many, including the principal architects of the early testing movement in the United States, as heritable and relatively immutable. Although debate continues as to the degree to which intelligence is either heritable or immutable, the concept of intelligence and the distinction between standardized tests of ability and standardized tests of achievement are tightly interwoven themes in American education.

Finally, opportunity to learn (or OTL) refers to the resources available to students, most often in the classroom setting, that facilitate their acquisition of knowledge or skills. This is a relatively simple definition of OTL, a term that is contested in the policy and educational research literature (for examples of other definitions, see Taiman 1993). Nevertheless, OTL has been implicitly invoked in testing since the earliest standardized tests, as described below. The concept was first formally defined in psychology by Carroll (1963), who equated OTL with time allowed for learning. Definitions proposed by Sorensen & Hallinan (1977) and Sorensen (1989) are more expansive and include teacher pedagogy, ability and effort, instructional pacing, and curriculum.

OTL has been a critical point of contention in standardized testing in American education. As discussed below, it has been invoked by both supporters and critics of standardized testing and has been deployed in the public sphere and in the courts. OTL is also key to our understanding of the role of standardized testing in social stratification; the extent to which testing functions to reduce or expand the educational opportunities available to disadvantaged youth and the variance in educational outcomes across groups of students hinges on the distribution of OTL. Depending on how learning actually occurs, expanding OTL may result in reductions in educational inequality (Kilgore & Pendleton

1993) or have the perverse effect of increasing inequalities (Sorensen & Hallinan 1977).

THE EVOLUTION OF STANDARDIZED TESTING IN AMERICAN EDUCATION

In 1916, Lewis Terman modified a test developed in France by Alfred Binet to identify children in need of special educational services owing to mild retardation or learning disability. Fueled by the eugenics movement, Terman wished to sort all children by their cognitive ability. The test was first applied to the task of classifying students for instructional purposes in Oakland, California, in 1917 (U.S. Congr. Off. Technol. Assess. 1992) and spread rapidly across the United States thereafter. By 1925, around 90% of elementary schools and 65% of urban high schools grouped students by ability, many on the basis of intelligence tests (U.S. Congr. Off. Technol. Assess. 1992). The testing movement did not drive the decision to group students by ability for instruction; ability grouping was first formally proposed by Charles W. Eliot, president of Harvard University, in the 1890s as a means of making instruction more efficient. Instead, the cult of managerial efficiency that was spreading during that time may have contributed to the diffusion of both ability grouping and testing (Chapman 1988).

A number of features of the United States at the beginning of the twentieth century contributed to the spread of standardized educational testing. On the supply side, the development of correlational methods in statistics combined with advances in measurement spawned a body of psychological work on the measurement of cognitive ability. The initiation of a group-based pen and paper version of the Stanford-Binet test in the guise of the Army Alpha Test (the first multiple-choice test) dramatically reduced the cost of measuring an individual's cognitive ability (prior to the development of the alpha test in World War I, the Stanford-Binet was individually administered). On the demand side, extraordinary increases in immigration coupled with a rise in the share of youth attending school and an increased press for efficiency led schools to seek out ways of improving their instruction for an increasingly diverse pool of students (Fass 1980, U.S. Congr. Off. Technol. Assess. 1992). Testing was also embraced by the Progressive movement as a means of tailoring instruction to the skills, tastes, and instructional needs of individual students and of making educational decisions based on student merit rather than race or social class (Ackerman 1995, Fass 1980).

At the same time that ability tests were gaining in popularity, achievement tests were also evolving. Whereas tests of ability claimed (or aimed) to be free of the influence of culture, class, and curriculum, achievement tests did not. Achievement tests sought to assess command over specific curricular content and made no claims as to the heritability or stability of cognitive ability. To the contrary, architects of achievement tests sought to measure student knowledge and student learning.

E.L. Thorndike and his students at Columbia University were among the first to produce standardized tests of achievement in 1908. By 1917, more than 200 standardized tests of achievement were in circulation (Chapman 1988). In 1929, E.F. Lindquist developed the first major voluntary statewide achievement testing program, the Iowa Test of Educational Development. Used to diagnose and monitor educational progress, the Iowa test established the feasibility of large-scale statewide testing programs. According to Lemann (1999), the Iowa tests were "part of a broad, optimistic, democratic view of education. Lindquist wanted to educate more students, not fewer, and to use tests to further the goal" (p. 25).

The evolution of standardized testing in the United States illustrates three fundamental contradictions around testing that persist today. The first contradiction is between testing as a means of social redistribution and as a means of social reproduction. As Cronbach (1975) notes, "[p]roponents of testing, from Thomas Jefferson onward, have wanted to open doors

for the talented poor, in a system in which doors are often opened by parental wealth and status" (p. 1). Take, as an example, the emergence of the SAT. The two men arguably most responsible for the evolution of the SAT were Henry Chauncey, the first director of the Educational Testing Service, and James Bryan Conant, president of Harvard University from 1933 to 1953. Chauncey wanted to improve the measurement of human abilities, to create a census of the abilities of the population (Lemann 1999), and to aid in the development of each individual based on their interests and capabilities. Conant's objective was to "unseat the Episcopacy" at Harvard University, replacing the children of the privileged with more academically and intellectually elite men from across the country (and from across class lines). Together, Chauncey and Conant approached Carl Brigham, a principle in the IQ testing movement, and modified his SAT to serve as one of several components of a scholarship screening program for Harvard University. Chauncey and Conant did not advocate for universal access to higher education. However, both men believed that the distribution of academic aptitude was distinct from the distribution of privilege [we follow Cronbach & Snow (1977) in conceiving of aptitude as whatever facilitates learning or performance in a particular domain (cited in Messick 1981)]. They hoped to use the SAT to locate young men of high academic aptitude who had not enjoyed the learning opportunities of more privileged youth and thus might appear unqualified on the basis of their academic achievements.

Critics of the SAT, however, claimed (and continue to claim) that the SAT was used by elite universities to limit the access of historically excluded groups to higher education (Alon & Tienda 2007, Soares 2007) and that "gatekeeping tests such as the SAT are an important linchpin in a system dominated by social class, sustaining a highly stratified system of higher education" (Sacks 2007). Others go even further, alleging that the SAT is nothing more than a test of the affluence of a student's parents (see authors cited in Zwick 2004a) and has the effect of restricting college opportunities for low-income students (Crouse & Trusheim 1988). Finally, critics assert that private and expensive SAT preparation courses increase scores for those who can afford to take them, thus exacerbating the already pronounced socioeconomic inequalities in SAT scores (C. Buchmann, V.J. Roscigno, D.J. Condron, unpublished manuscript).

The second major contradiction is between competing notions of equality of opportunity. Under the common schools movement that helped shape public education between the 1850s and early 1900s, the operative definition of equality was sameness; all students should experience the same curriculum in schools of the same quality with similar if not identical levels of resources. In the early twentieth century, however, the classical college preparatory curriculum of the high school was challenged by the National Education Association as irrelevant to the majority of students whose occupational trajectories did not include college (Coleman 1990). Instead, some argued that equality of opportunity required a differentiated curriculum that would allow students to pursue varied occupational outcomes, whereas progressives argued for a curriculum more closely aligned with the skills and abilities of individual students. The implicit definition of equality of opportunity here was one of differentiation, not commonality.

If differentiation were to be achieved, standardized tests appeared well-suited to the task of allocating educational opportunities both fairly and efficiently (U.S. Congr. Off. Technol. Assess. 1992). At the postsecondary level, E.F. Lindquist argued in the 1950s that the SAT failed to differentiate satisfactorily among prospective students. He proposed a college entrance exam that would be based on achievement rather than ability and aptitude, that would be aligned more closely to the secondary school curriculum, and that would help less prestigious colleges both admit and place students in courses. The outcome of his efforts, the ACT, was first administered in November 1959. It remains the chief competitor to the SAT and

is now accepted as an alternative to the SAT by all colleges and universities that require an admissions test.

The third fundamental contradiction surrounding standardized testing is its effect on curricular content: Do and should tests drive curricular and pedagogical decisions? When Harvard President Charles Eliot first proposed a cooperative system of exams for college admissions in 1890, he had two goals in mind—to assess the academic readiness of students and the adequacy of the high schools from which they graduated. In fact, Eliot and others who ultimately joined together as the College Entrance Examination Board in 1900 sought to use the college entrance exam to leverage changes in the curriculum of secondary schools. By the end of World War I, they had apparently been successful, judging by the complaints of some teachers and administrators who found that the exams dominated curriculum and pedagogy, constraining their ability to offer the curriculum they would like to offer to students (U.S. Congr. Off. Technol. Assess. 1992). Some researchers and teachers continue to argue that the SAT has an adverse impact on the high school curriculum today (Shepard 1993). As we discuss in the section entitled "Contemporary Uses of Tests," below, this third contradiction is a major factor in the battles surrounding the No Child Left Behind Act (NCLB).

SHOULD WE SHOOT THE MESSENGER?

Underlying some of the critiques of standardized testing is the charge that it inadequately measures academic achievement, cognitive ability, or academic aptitude. According to Jencks (1998), for example, "[m]any Americans discount racial differences in test performance on the grounds that all cognitive tests are either racially or culturally biased" (p. 55). Although this is certainly so for some, public opinion polls generally show that the majority of parents do not believe that standardized tests are biased. A 1998 survey, for example, found that only 28% of African American parents felt that standardized tests are culturally biased against black students (Henry 1998).

In the sections below, we review three key properties of tests—validity, reliability, and norming—that bear directly on their utility and fairness. We then return to a discussion of OTL and on the relationship between OTL and the validity and fairness of standardized tests.

Validity

The concept of test validity has evolved substantially from the turn of the century when it was simply defined in terms of the accuracy of a test's estimate of a student's ability or achievement. Today, researchers suggest that validity is more of a process than a product. As Kane (2001) succinctly puts it, "[v]alidity is concerned with the clarification and justification of the intended interpretations and uses of observed scores."

Although the concept of validity has become more nuanced and holistic over time, we discuss three complementary dimensions of validity for expository purposes. Prior to a landmark article by Cronbach & Meehl (1955), validity was established primarily by reference to some criterion such as another test, a performance, or some other observable outcome. Criterion validity remains an important element of validity today, but only in situations that include a "well-defined and demonstrably valid criterion measure" (Kane 2001). Such situations are exceedingly rare in the field of education.

Construct validity was originally proposed as a complement to criterion validity that aimed to incorporate better the processes of measurement and theory building (Cronbach & Meehl 1955). Under construct validity, "[i]f the observations are consistent with the theory, the validity of the theory and of the measurement procedures used to estimate the constructs defined by the theory are both supported. If the observations are not consistent with the theory, some part of the network would be rejected . . ." (Kane 2001). Construct validity changed the focus of validation from validating a test instrument to validating a proposed interpretation of a test instrument.

Finally, the concept of consequential validity has stimulated substantial debate in the psychometric literature. Consequential validity is in many ways a natural extension of construct validity, expanding the concept beyond underlying theory and constructs to include the consequences of using a particular assessment for a particular purpose. A test score that accurately measures what it seeks to measure and yet is used to deny students opportunities from which they would benefit may lack consequential validity.

Reliability

Like validity, reliability is a property of a test score used for a particular purpose in a particular context. Whereas validity at least in part speaks to the degree to which the expected value of a measure approximates the attribute it seeks to estimate or predict, reliability is concerned solely with errors in measurement, generally estimated by the stability of a measure across replications. The theoretical universe of those replications is defined by the investigator on the basis of conditions of testing or scoring that the investigator believes should be random with respect to the examinee's score (Brennan 2001; for a thorough and more technical treatment of reliability, see Haertel 2006).

Bias

Bias is a measure of the degree to which the expected value of a test score deviates from the true value of the object the test seeks to measure and is therefore closely related to construct validity. In the literature on testing, however, bias rarely applies to a population difference in expected and true values. Instead, bias generally refers to a difference between the expected and true means of some characteristic for one subpopulation (e.g., African Americans or women) relative to another (e.g., whites or men). Many of the early proponents of IQ testing, including Binet, were sensitive to issues of test bias because they sought a culture-free measure of cognitive ability (Camilli & Shepard 1994).

Group differences on an assessment are neither necessary nor sufficient to indicate that an assessment is biased. Take as an example the difference in mean SAT math scores for children with at least one parent with a graduate degree and for those with no parent who attended college (College Board 2006). The difference—93 points, or roughly eight-tenths of a standard deviation—may accurately reflect differences in the mathematics aptitude of these two groups that are influenced by any number of factors, including home and school environments, OTL, and preferences for taking advantage of those opportunities when they are available. On the other hand, the difference might also reflect group differences in verbal ability or ancillary knowledge relevant to answering some of the test questions or other factors not directly related to mathematics aptitude but correlated with parental education. If the latter is true, the estimated difference is upwardly biased.

Cole (1981) suggests two different approaches to the study of bias, one based on the idea of criterion validity and the other on construct validity. From the perspective of criterion validity, a test score is biased if its relationship with the criterion varies across subgroups. Criterion-related bias would be reflected by a regression of the criterion on the test score that revealed different slopes for the test score or different intercepts across groups. Cole (1981) notes that "a consistent finding has been that the use of a single regression equation based on combined data from black and white groups results in the overprediction of performance by black students." This is certainly true for the SAT, which overpredicts freshman GPA for African American students by around a tenth of a grade point relative to white students (Young 2004).

Another important aspect of bias related to construct validity—item bias—inheres in the interrelationships among test items across groups. An unbiased test is one in which interrelationships are uniform across groups. Item bias has attracted a great deal of attention from test designers and critics alike. The most transparent analysis of item bias is a review of items

by experts to ferret out those prompts that may favor or disfavor certain groups. Critics of testing argue that test items are often, if not routinely, biased against African Americans, Latinos, women, and/or socioeconomically disadvantaged test takers (FairTest 2007). The Educational Testing Service has developed increasingly sophisticated approaches to expert review for item bias since the 1960s (Educational Testing Service 2003), but past empirical research suggests that "expert judges did no better than chance in predicting which items would be relatively more difficult for black examinees" (Camilli & Shepard 1994).

Once items have been administered to a sample of examinees, the next more technical stage of bias analysis, the analysis of differential item fit (or DIF), can begin. DIF analysis compares the fit of each test item for members of different groups conditional on their total test score. Items that differ in their difficulty or ability to discriminate across levels of student aptitude across groups are considered biased and eliminated from the test score and future test occasions. In the case of the SAT, potential items are piloted in the unscored section of the SAT given to examinees annually and carefully analyzed for DIF. For a more comprehensive review of technical issues related to DIF and the continuing evolution of DIF techniques, see Camilli (2006).

DIF analysis is firmly rooted in the tradition of construct validity and thus subject to the hazards of lacking a clear criterion. DIF analyses are only effective at ferreting out item bias if the overall scores are not themselves severely biased, a limitation of construct validity that some critics have highlighted to argue that the SAT is biased against African Americans. For example, Rosner (2003) characterizes the SAT as a test composed of white preference items because items that a higher share of white students than black students answer correctly are included in the test, whereas those that a higher share of black students answer correctly are excluded. Freedle (2003), in contrast, proposes a revised SAT score (R-SAT) that corrects what he argues are cultural and statistical biases in the SAT by scoring only the harder half of the items on the verbal assessment. He finds that, for more difficult items on the SAT, African American students with relatively low estimated scores performed better than white students with relatively low estimated scores, leading to a reduction in the difference in average verbal achievement scores of one-third of a standard deviation. Freedle's methods and results, however, are impugned by Dorans (2004), who argues that the R-SAT, properly estimated, would result in very small changes to the black-white gap in SAT scores.

Norm and Criterion Referencing

Unlike measures of weight or height, test scores do not have their own generally agreed upon metrics; they are measures of unobservable characteristics. To interpret test scores, therefore, they must be associated with a set of reference points. One way of making scores meaningful—norm referencing—is to use them to locate an examinee on the distribution of scores of all examinees. The meaning of norm-referenced scores is bound to the norming population, the group of examinees on whose scores the rank ordering is based. Norm-referenced tests are commonly used in education for a variety of purposes, including comparing students, districts, states, and nations to each other and (in some jurisdictions) making grade retention and course placement decisions.

Although norm referencing facilitates comparing each examinee to a population of examinees, criterion referencing supports comparisons of examinees to levels of subject matter knowledge or proficiency. Glaser (1963) first proposed criterion-referenced measures as a more valid way of assessing student competence and achievement, arguing that we should compare students' achievement to the level of competence we wish them to achieve.

Criterion-referenced tests (CRTs) have made several useful contributions to testing. First, CRTs heighten the importance of articulating the intended outcomes of instruction. In fact, CRTs enabled the standards-based

reform movement; without a clear way of linking assessments to the standards intended to guide instruction, there was little way of knowing whether instructional objectives were being met. Second, CRTs should in principle reduce the problems associated with teaching to the test. CRTs, to the extent that they accurately measure the objectives laid out in instructional standards, are tests worth teaching to (Linn 1993).

Some critics mischaracterize norm- and criterion-referenced tests as antithetical. In fact, although a norm-referenced test need not have a criterion-referenced interpretation, any CRT can be normed (Linn 1994). The distinction between norm- and criterion-referenced tests is often overstated because the criteria behind CRTs are often inherently normative (Angoff 1974). Take, for example, criterion-referenced minimum competency exams. What counts as minimally competent varies across place and time for both economic and sociological reasons. There is nothing absolute about the threshold at which one defines minimally competent (or, in the parlance of NCBL, proficient).

Opportunity to Learn and Issues of Measurement

Many critics have emphasized the pernicious effects of differences in OTL in their arguments against the use of standardized tests for high-stakes decisions such as grade retention, high school graduation, and college admissions. Implicit in this criticism, at least in the case of college admissions, is a preference for measuring students' academic potential rather than their realized achievement. Although academic potential was precisely what the SAT set out to measure in 1926, even the makers of the SAT have long acknowledged that the test is a measure of "students' *developed* analytical skills and abilities," something far closer to an achieved than an ascribed attribute (Caperton 2006, emphasis added).

Although OTL is an extremely important substantive concern and, from the perspective of social stratification, may be the most important issue in testing, it is also a measurement issue. If one wishes to estimate the counterfactual outcome of what achievement would have been under greater or lesser levels of OTL, one's estimates based on realized academic achievement are certainly biased owing to the correlation between OTL and test performance. In the case of norm-referenced tests, the measurement problem can be framed as one of having the wrong norming population. One should norm not on the entire population of test takers, but instead on the subpopulation of test takers experiencing identical levels of OTL. In fact, some scholars have advocated for a form of OTL-based renorming of the SAT in a regression framework (e.g., Carnevale & Haghighat 1998, Studley 2006). A larger body of research argues for the utility of renorming student scores for the purposes of test-based school accountability in the form of value added models of student achievement (Coleman et al. 1997, Meyer 1997; cf. Rubin et al. 2004).

CONTEMPORARY USES OF TESTS

Standardized testing permeates contemporary formal education in the United States. Standardized tests are now used for descriptive and research purposes, diagnostic purposes, and gatekeeping. Tests are also increasingly used to hold educators and administrators accountable for the academic achievements of their students. Below we describe some of the contemporary uses of testing in education. We conclude this section by describing threats to the validity of tests for these purposes, some contemporary misuses of tests, and some alternatives to standardized testing proposed by researchers and practitioners.

Testing for Description and Research

With funding from the federal government, researchers have been administering standardized tests to large samples of American youth since 1960, when Project TALENT assessed over

400,000 high school students (Flanagan et al. 1964). The most widely recognized and consequential of these studies may be the *Equality of Educational Opportunity Study*, which led to the Coleman Report (Coleman et al. 1966). Funded by Congress in 1964 to document the relationship between educational resources and academic achievement differences between African American and white youth, Coleman and colleagues instead found that school resources account for relatively little of the variation in student academic achievement. In fact, schools themselves appeared to account for only around one-fifth of that variation.

To monitor the academic progress of American students across cohorts, the U.S. Department of Education has periodically assessed students' achievement in mathematics, reading, science, writing, and other subjects since 1969 under the NAEP. NAEP estimates achievement for populations or subpopulations of students (e.g., fourth-graders or African American students) at the national, regional, and state levels, but does not estimate the achievement levels of individual students.

At the state level, most states have had extensive student testing policies for several years. State tests are typically aligned with state standards for what teachers should teach and what students should learn (American Federation of Teachers 2006). Putting aside their use for student and school accountability purposes (described below), these tests are fundamentally designed to describe the academic achievements of students and their educational institutions.

Efforts to describe and analyze international differences in student achievement are facilitated by a number of major cross-national testing programs. The Trends in International Mathematics and Science Study (TIMSS) study, for example, has assessed trends in students' mathematics and science achievement every four years since 1995. More than 60 countries are participating in the 2007 round of TIMMS. Likewise, the Program for International Student Assessment tests 15-year-olds' reading, mathematics, and science literacy in over 60 countries (OECD 2006).

With the exception of Project TALENT, each of the testing programs described above is cross-sectional, and many include limited information about students' backgrounds or other measures of social context. Since 1960, the U.S. Department of Education has carried out several longitudinal studies of American students that have included such measures in addition to achievement tests in multiple subjects and at multiple points in students' schooling careers. Different studies have followed cohorts of students starting at the primary and preprimary levels, middle school, high school, and college. A number of other longitudinal studies of young people outside the U.S. Department of Education have included achievement or ability tests as well. Each of these longitudinal studies allows investigators to ask important questions about social, economic, and demographic variation in test scores and to model the predictors and consequences of academic achievement.

Testing for Diagnostic Purposes

Standardized diagnostic tests are routinely used to determine students' eligibility for educational support and benefits under a variety of state and federal programs, including benefits prescribed by the federal Individuals with Disabilities Education Act (IDEA) and its state variants and by federal and state bilingual education programs for English-language learners (ELL). In some cases—for example, with respect to determining eligibility for accommodations under IDEA—testing is just one of several criteria used in a broader program of evaluation and assessment. In other cases—for example, in determining students' status as ELL—diagnoses rely much more heavily on the results of standardized tests.

Testing for Gatekeeping

Frequently motivated by a desire to end social promotion—the practice of advancing

students through the educational system despite their not having mastered grade-level specific curricular material—legislators from both sides of the political aisle have supported the use of testing for making decisions about whether students move from one grade in school to the next (Heubert & Hauser 1999). At least ten states (Snyder et al. 2006, table 156 therein) and one in three school districts (American Federation of Teachers 1998) used standardized tests for the purpose of making promotion or retention decisions at the end of the twentieth century.

Since the middle of the twentieth century, testing has come to play an increasingly important role in students' efforts to complete secondary schooling. Originally designed to provide World War II service members with an alternative route to college admission, the GED has a long history as a second-chance test-based credential. By the early 2000s, more than one million students took the GED annually (General Educational Development Testing Service 2003). The American Council on Education's (ACE) GED Testing Service produces and administers the tests, and its Commission on Educational Credit and Credentials determines the minimum passing standards. State educational agencies often apply higher thresholds. In 2002, ACE realigned GED passing scores to a level sufficient to ensure that 40% of graduating seniors would not pass (General Educational Development Testing Service 2003).

Increasingly, even students pursuing their high school diplomas through traditional means must pass standardized achievement tests. The number of states with high school exit exams rose gradually from zero for the class of 1978 to 22 for the class of 2006 (Center on Education Policy 2003, Warren & Kulick 2007). Because some relatively populous states have enacted such policies (e.g., California, New York, Texas, Florida, and Ohio), about two-thirds of all high school students in the graduating class of 2006 faced such a requirement. There is wide state-to-state variability with respect to the subjects covered on states' high school exit examinations, the difficulty level of exit examinations, policies concerning the testing of students with disabilities or who are ELL, the grade level at which the examinations are first administered, and passing thresholds.

Testing has also played an important role in admission to postsecondary and professional schooling since the middle of the twentieth century. Whereas only a few hundred postsecondary institutions considered scores on the SAT or the ACT for purposes of admissions to undergraduate programs in the 1950s, almost all four-year colleges and universities considered SAT or ACT scores by the end of the twentieth century. In recent decades, graduate and professional schools have also increasingly relied on standardized tests for admissions decisions. Finally, a number of professions rely on standardized testing for licensure and certification. For example, at least 29 states required prospective elementary or secondary school teachers to pass basic skills tests as a prerequisite for initial certification in 2006, and at least 35 required prospective teachers to pass subject-area tests (Natl. Assoc. State Dir. Teach. Educ. Certif. 2006).

Testing and School Accountability

Although many states had (and continue to have) their own test-based district and school accountability programs, the 2001 NCLB added a federal mandate that all states implement accountability systems that cover all public schools. NCLB allows states considerable flexibility in designing their accountability systems, but each state must base their system on annual testing for all students in grades 3–8 and one high school grade. Tests must be aligned with well-defined standards in reading, mathematics, and science. Schools and districts that fail to make adequate yearly progress (AYP) toward statewide proficiency goals are subject to restructuring if they do not improve or take sufficient corrective action. The NCLB currently requires that AYP—which is centrally based on the percentage of students scoring at or above targeted thresholds on achievement tests—be attained for all sufficiently large

subgroups of students defined by poverty status, race/ethnicity, disability status, and limited English proficiency status.

Unlike primary, middle, and secondary schools, colleges and universities have not been subject to testing requirements for accountability purposes in the past, nor has learning in the college been subject to as much scholarly scrutiny. The available evidence, based largely on standardized test scores for students attending four-year colleges, suggests that student academic achievement over the undergraduate years increases by around two-thirds of a standard deviation in English, one-third of a standard deviation in math, and one-half of a standard deviation in science, net of several student characteristics including secondary school academic achievement (Pascarella & Terenzini 2005). Much of this growth appears to take place during the first two years of college. There is little evidence that achievement growth varies across colleges in a substantively meaningful way.

Today, colleges and universities are under increasing pressure to demonstrate growth in students' academic achievement. Spurred into action by calls for postsecondary accountability issued by the Spellings Commission (Secr. Educ. Comm. Future High. Educ. 2006), the American Association of State Colleges and Universities and the National Association of State Universities and Land-Grant Colleges have established a Voluntary System of Accountability (VSA). In addition to providing information on college graduation rates and costs of attendance, the VSA will also include a measure of student learning outcomes using one of three approved assessment instruments (see **http://www.voluntarysystem.org**). It is far too early to tell how many institutions will participate in the VSA in general and more specifically in the pilot project of including changes in standardized test scores among their learning outcomes (colleges and universities were invited to begin registering for the VSA in December 2007). However, if the pressure for postsecondary accountability persists, we expect to see an increase in testing of college students and, as a result, an increase in the research base on student learning in higher education.

Threats to Validity

Contemporary uses of standardized testing are frequently invalid with respect to construct and/or consequential validity. Recall that construct validity begins with the articulation of both theory and measurement. What is the theory underlying the use of testing for grade promotion and retention decisions? Is it that students whose scores fall below some minimum threshold would not enjoy any benefit from moving on to the next grade? Or is it that an alternative treatment (repeating a grade) would lead to a greater increase in their knowledge or skills? One could ask the same questions of tests for college and professional school admissions, licensure, or school accountability.

A related issue touches on both construct validity and reliability: What is it that contemporary standardized tests seek to assess and what else drives variation in test scores? For example, in the case of grade promotion decisions, do we wish to isolate the contribution of student effort and cognitive ability to variation in test scores from contributions of teachers, schools, and neighborhoods? Or do we want to retain students in a grade regardless of the source of their test failure? Likewise, when we recruit and select students for college, do we wish to do so based on their observed or potential performance?

The consequential validity of contemporary uses of standardized tests is also frequently questionable. For example, there is substantial evidence that tests are invalid for making decisions about grade retention or for awarding high school diplomas. For standardized tests to enjoy consequential validity for such purposes, students' treatments as a result of decisions based on standardized tests must lead to better outcomes than the treatment that would have been administered in the absence of the test. Research strongly suggests that grade retention does more harm than good, increasing

the risk of dropout without any demonstrable improvement on academic achievement (Hong & Raudenbush 2005). Likewise, state high school exit examinations deny diplomas to a small share of otherwise eligible seniors without any clear benefit in terms of academic achievement (Grodsky et al. 2008), postsecondary participation, or labor market outcomes (Warren et al. 2008).

Inequality in OTL underlies many problems of construct and consequential validity and directly relates construct validity to social stratification. Variation in OTL may drive a substantial share of both test score variance and between-group differences in test score means. This does not necessarily undermine the construct validity of assessments; to the extent that assessments are designed to measure subject matter knowledge or academic achievement, variation in OTL may contribute to valid variation in test scores. However, if policy makers or decision makers wish to isolate the contribution of student effort and/or innate ability from contributions of environmental factors such as OTL, variation in OTL introduces substantial error into estimates of these quantities based on scores on standardized tests. Variation in OTL is a more serious challenge to consequential validity. If students are assigned to treatments (grade promotion, high school graduation) based not on their own likely outcomes but instead on OTL, we can virtually guarantee that those students most poorly served by the educational system will disproportionately be those further punished through grade retention and the denial of high school credentials, college admissions, financial aid, and other resources.

Misuse of Tests

In a 1999 report the National Research Council noted that there is often a gap between policy makers' expectations of testing and the technical capacities of the tests that they employ. "One of the most common reasons for this gap," they argued, "is that policymakers . . . often decide to use existing tests for purposes for which they were neither intended nor adequately validated" (Heubert & Hauser 1999, p. 31). One example of this was the Chicago Public Schools use in the 1990s of the Iowa Test of Basic Skills—a nationally norm-referenced achievement test—to identify locally underperforming students and schools, even though it was never designed for such purposes. Another familiar example was Secretary of Education Terrell Bell's mid-1980s "wall chart," which ranked states based on students' mean SAT and ACT scores. Neither the SAT nor the ACT were designed to measure students' mastery of curricular materials, and, in any case, those tests are only taken by a selective subset of students (with the degree of selectivity varying across states).

CONCLUSION: REVISITING THE ROLE OF TESTING IN SOCIAL STRATIFICATION

The most fundamental contradiction in standardized testing, from a stratification perspective, is that between social redistribution and social reproduction. Empirical patterns of measures of cognitive ability and academic achievement are, at first glance, more consistent with the reproduction than the redistribution perspective. After all, if tests enabled redistribution we should not observe any association between social origins and test scores. And if we observed such a relationship in the past, we should see it decline precipitously and uniformly over time.

Underlying the logic that standardized tests work to redistribute social position is a belief that standardized tests themselves exert some causal force on outcomes independent of the underlying attributes those tests seek to measure. We question this belief. If tests enjoy high levels of construct validity and reliability, as is certainly the case for almost all well-vetted standardized assessments in use today, the additional contribution of the scores themselves (above and beyond the cognitive ability or academic achievement they represent) is likely to be quite modest. This is not to say that standardized tests are perfect instruments;

they are not. We agree with the American Psychological Association, the American Educational Research Association, and most test makers (including the Educational Testing Service): High-stakes decisions should never be made based on tests alone. That they are used for such purposes is more a testament to a poverty of policy than to flaws in measurement. Test scores for the most part reflect differences in qualities among students; policy makers, educators, and gatekeepers decide how to respond to those differences.

Some critics of standardized tests such as the SAT seek to impeach the tests based on correlations between test scores and cognitive ability or IQ on the one hand and correlations between test scores and measures of SES on the other. Given the state of social stratification in the United States, however, we find these correlations neither surprising nor (from a measurement perspective) problematic. First, to the extent that smarter students learn more or at a faster rate, we should expect cognitive ability and academic achievement to be positively related to one another. Second, nobody disputes the assertion that more advantaged parents seek to pass advantages along to their children through school, both by engaging them in school-like activities in the home from a young age and by seeking to ensure that they attend the best schools and get the best teachers. Notwithstanding the question of whether or not cognitive ability is to any degree heritable, we would be surprised to find a test of academic achievement in which more advantaged students failed to outscore their less advantaged peers. Our society is stratified along the lines of race and SES; standardized test scores reflect that fact. To try to hide this truth by designing a test that shows no difference in the achievement of those historically advantaged and those consistently denied OTL would be both disingenuous and regressive.

On the other hand, variation within groups is generally much greater than variation between groups. It would be too simplistic to focus only on differences in mean scores, as Hedges & Nowell (1999) demonstrate. Furthermore, without carefully defining equality it is not obvious what a just distribution of standardized test scores would look like. Uniformity in mean scores across groups would require equalizing OTL both within and without the classroom and extending the concept to parenting practices, tastes and preferences of parents and their children, and even birth weight. Given the extent to which things such as parenting practices, habits, and tastes lie beyond the pale of public policy, it seems more reasonable to focus on factors we can more readily influence: the distribution of OTL in the schools.

OTL standards (and, before them, school delivery standards) are not new in American educational policy. The idea of measuring school inputs in the form of funding, teacher credentials, instructional resources, and other materials has been around for a long time. In their review of past efforts to increase educational quality by carefully monitoring and controlling inputs, Elmore & Fuhrman (1993) conclude that OTL standards hold little promise for affecting change in the educational system. Likewise, Porter (1993, 1995) cautions against using OTL or school delivery standards for school accountability because of uncertainty around what exactly should be included in OTL standards, a reluctance to be overly prescriptive, and concerns over the substantial technical issues involved with effectively measuring OTL. As a practical matter, McDonnell (1995) suggests that states are unlikely to articulate OTL standards out of fear of litigation.

We take the concerns raised by Porter (1993) seriously, but at the same time agree with Porter that "[i]t would be unfair to hold students to new and more rigorous standards of achievement without holding schools accountable for giving students a fair chance at meeting those standards." Porter does not believe that this requires holding schools accountable to delivery standards. We disagree with Porter but offer an alternative: Suspend student accountability at schools that fail to meet basic OTL standards. Retaining a child in a grade only to expose her to the same ineffective instruction that led her to score under a test benchmark or denying a

young man his high school diploma as a result of denying him reasonable educational services for the past 4–12 years of his life is unjust. Using the cumulative number of students retained or denied diplomas as a means of holding schools accountable is both unjust and ineffective.

Ideally, schools should adjudicate among students based on a combination of students' effort and academic aptitude, not based on the quality of instruction to which they have access. High-stakes testing often fails to meet this standard by virtue of holding students accountable for learning material to which they have been inadequately exposed, if at all. To prevent additional harm from being done to students denied adequate OTL, we argue that we need to have some means of measuring OTL. In our view, this is the most critical issue underlying the use of standardized tests for high-stakes decisions in American education and the role of standardized tests in social stratification.

Although we agree with many critics who contend that there is far too much standardized testing in contemporary American education, we do not believe that test scores are the main problem. The real issue is the stratified distribution of knowledge and skills reflected in the scores that students earn on standardized tests. In the context of formal schooling, we believe that OTL is the linchpin of social stratification. On the one hand, given the glaring inequalities students experience from birth onward beyond the classroom, we think it unlikely that anything short of a heroically progressive redistribution of OTL would result in the amelioration of the test score differences we observe across race/ethnicity and socioeconomic origins. On the other hand, we also remain somewhat uncertain about the effect that increasing OTL across the board would have on stratification in student learning: It could reduce stratification (Kilgore & Pendleton 1993) or further exacerbate it (Sorensen & Hallinan 1977).

We encourage researchers to develop further the concept of OTL and our understanding of its role in the stratification of educational achievement and attainment. We also endorse efforts to restandardize scores on the SAT advocated by Carnevale & Haghighat (1998) and Studley (2006) to adjust for differences in OTL and believe that their work may have implications for primary and secondary education. Nonetheless, we caution against relying on approaches such as these as remedies for stratification in knowledge and skills. While we continue to accommodate those denied adequate OTL, we should remain committed to remedying the inequalities that produce gaps in standardized test scores rather than using statistical procedures to remedy the disservice we do to many of our youth.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Ackerman M. 1995. Mental testing and the expansion of educational opportunity. *Hist. Educ. Q.* 35:279–300

Alon S, Tienda M. 2007. Shifting meritocracy in higher education. *Am. Sociol. Rev.* 72:487–511

American Federation of Teachers. 1998. *Passing on Failure: District Promotion Policies and Practices*. Washington, DC: Am. Fed. Teach.

American Federation of Teachers. 2006. *Smart Testing: Let's Get It Right*. Washington, DC: Am. Fed. Teach.

Angoff WH. 1974. Criterion referencing, norm referencing, and the SAT. *Coll. Board Rev.* 92:3–5, 21

Boocock SS. 1966. Toward a sociology of learning: a selective review of existing research. *Sociol. Educ.* 39:1–45

Brennan RL. 2001. An essay on the history and future of reliability from the perspective of replications. *J. Educ. Meas.* 38:295–317

Buchmann C, DiPrete TA, McDaniel A. 2008 . Gender inequalities in education. *Annu. Rev. Sociol.* 34:319–37

Camara WJ, Schmidt AE. 1999. Group differences in standardized testing and social stratification. *Coll. Board Rep. No. 99-5*, New York, College Entrance Examination Board

Camilli G. 2006. Test fairness. In *Educational Measurement*, ed. RL Brennan, pp. 221–56. Westport: Praeger

Camilli G, Shepard LA. 1994. *Methods for Identifying Biased Test Items*. Thousand Oaks: Sage

Caperton G. 2006. Doing what is important in education. See Zwick 2004b, pp. 33–36

Carnevale AP, Haghighat E. 1998. Selecting the strivers: a report on the preliminary results of the ETS "Educational Strivers" study. In *Hopwood, Bakke and Beyond: Diversity on Our Nation's Campuses*, ed. DL Bakst, pp. 122–28. Washington, DC: Am. Assoc. Coll. Regist. Admiss. Off.

Carroll JB. 1963. A model of school learning. *Teach. Coll. Rec.* 64:723–33

Carroll JB. 1997. Psychometrics, intelligence, and public perception. *Intelligence* 24:25–52

Center on Education Policy. 2003. *State High School Exit Exams: Put to the Test*. Washington, DC: Cent. Educ. Policy

Chapman PD. 1988. *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890–1930*. New York: N. Y. Univ. Press

Cole NS. 1981. Bias in testing. *Am. Psychol.* 36:1067–77

Coleman JS. 1990. *Equality and Achievement in Education*. Boulder: Westview

Coleman JS, Campbell EQ, Hobson CJ, McPartland J, Mood A, et al. 1966. *Equality of Educational Opportunity*. Washington, DC: USGPO

Coleman JS, Schneider B, Plank S, Schiller KS, Shouse R, Wang H. 1997. *Redesigning American Education*. Boulder: Westview

College Board. 2006. *2006 College-Bound Seniors*. Princeton, NJ: College Board

Cronbach LJ. 1975. Five decades of public controversy over mental testing. *Am. Psychol.* 30:1–14

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52:281–302

Cronbach LJ, Snow RE. 1977. *Aptitudes and Instructional Methods*. New York: Wiley

Crouse J, Trusheim D. 1988. *The Case Against the SAT*. Chicago: Univ. Chicago Press

Dorans NJ. 2004. Freedle's Table 2: fact or fiction? *Harvard Educ. Rev.* 74:62–72

Educational Testing Service. 2003. *Fairness Review Guidelines*. Princeton, NJ: ETS

Elmore R, Fuhrman S. 1993. Opportunity to learn and the state role in education. In *The Debate on Opportunity to Learn Standards: Supporting Works*, ed. SL Traiman, pp. 73–102. Washington, DC: Natl. Gov. Assoc.

FairTest. 2007. *The SAT: Questions and Answers*. Cambridge, MA: FairTest. **http://www.fairtest.org/sat-questions-and-answers**

Fass PS. 1980. The IQ: a cultural and historical framework. *Am. J. Educ.* 88:431–58

Fischer CS, Hout M, Jankowski MS, Lucas SR, Swidler A, Voss K. 1996. *Inequality by Design: Cracking the Bell Curve Myth*. Princeton, NJ: Princeton Univ. Press

Flanagan JC, Davis FB, Daily JT, Shaycoft MF, Orr DB, et al. 1964. *The American High School Student*. Pittsburgh, PA: Project TALENT Office, Univ. Pittsburgh

Freedle RO. 2003. Correcting the SAT's ethnic and social class bias: a method for reestimating SAT scores. *Harvard Educ. Rev.* 73:1–43

Fryer RG Jr, Levitt SD. 2004. Understanding the black-white test score gap in the first two years of school. *Rev. Econ. Stat.* 86:447–64

Fryer RG Jr, Levitt SD. 2005. *The black-white test score gap through third grade*. NBER Work. Pap. 11049, Natl. Bur. Econ. Res.

General Educational Development Testing Service. 2003. *Who Passed the GED Tests? 2003 Statistical Reports*. Washington, DC: Am. Counc. Educ.

Glaser R. 1963. Instructional technology and the measurement of learning outcomes: some questions. *Am. Psychol.* 18:519–21

Gosa TL, Alexander KL. 2007. Family (dis)advantage and the educational prospects of better off African American youth: how race still matters. *Teach. Coll. Rec.* 109:285–321

Grodsky E, Kalogrides D, Siebens J. 2007. *Virtuous cycle or perverse openness? Race, ethnicity and social class reconsidered*. Presented at Annu. Meet. Am. Educ. Res. Assoc., Apr. 9–13, Chicago

Grodsky E, Warren JR, Kalogrides D. 2008. State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971–2004. *Educ. Policy.* In press

Haertel EH. 2006. Reliability. In *Educational Measurement*, ed. RL Brennan, pp. 65–110. Westport: Praeger

Hedges LV, Nowell A. 1995. Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science* 269:41–45

Hedges LV, Nowell A. 1999. Changes in the black-white gap in achievement test scores. *Sociol. Educ.* 72:111–35

Henry T. 1998. Black parents see no bias in school tests. *USA Today*, July 29, p. 1D

Herrnstein RJ, Murray C. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press

Heubert JP, Hauser RM, eds. 1999. *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, DC: Natl. Acad. Press

Hong G, Raudenbush SW. 2005. Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educ. Eval. Policy Anal.* 27:205–24

Jencks C. 1998. Racial bias in testing. In *The Black-White Test Score Gap*, ed. C Jencks, M Phillips, pp. 55–85. Washington, DC: Brookings Inst.

Jencks C, Phillips M. 1998. The black-white test score gap: an introduction. In *The Black-White Test Score Gap*, ed. C Jencks, M Phillips, pp. 1–51. Washington, DC: Brookings Inst.

Jensen A. 1969. How much can we boost IQ and scholastic achievement. *Harvard Educ. Rev.* 39:1–123

Kane MT. 2001. Current concerns in validity theory. *J. Educ. Meas.* 38:319–42

KewalRamani A, Gilbertson L, Fox MA, Provasnik S. 2007. Status and trends in the education of racial and ethnic minorities. *Rep. NCES 2007–039*, Washington, DC, Natl. Cent. Educ. Stat., Inst. Educ. Sci., U.S. Dep. Educ.

Kilgore SB, Pendleton WW. 1993. The organizational context of learning: framework for understanding the acquisition of knowledge. *Sociol. Educ.* 66:63–87

Kobrin JL, Sathy V, Shaw EJ. 2007. A historical view of subgroup performance differences on the SAT reasoning test. *Res. Rep. 2006–5*, New York, College Board, New York

Lemann N. 1999. *The Big Test: The Secret History of the American Meritocracy*. New York: Farrar, Straus and Giroux

Linn RL. 1993. Educational assessment: expanded expectations and challenges. *Educ. Eval. Policy Anal.* 15:1–16

Linn RL. 1994. Criterion-referenced measurement: a valuable perspective clouded by surplus meaning. *Educ. Meas. Issues Pract.* 13:12–14

Lubienski ST. 2002. A closer look at black-white mathematics gaps: intersections of race and SES in NAEP achievement and instructional practices data. *J. Negro Educ.* 71:269–87

McDonnell LM. 1995. Opportunity to learn as a research concept and a policy instrument. *Educ. Eval. Policy Anal.* 17:305–22

Messick S. 1981. The controversy over coaching: issues of effectiveness and equity. In *New Directions for Testing and Measurement: Issues in Testing—Coaching, Disclosure, and Ethnic Bias*, ed. BF Green, pp. 21–53. San Francisco: Jossey-Bass

Meyer RH. 1997. Value-added indicators of school performance: a primer. *Econ. Educ. Rev.* 16:283–301

Natl. Assoc. State Dir. Teach. Educ. Certif. 2006. *The NASDTEC Manual on the Preparation & Certification of Educational Personnel in the United States & Canada*. Whitinsville, MA: NASDTEC

Neal D. 2006. Why has black-white skill convergence stopped? In *Handbook of Economics of Education*, ed. EA Hanushek, F Welch, pp. 511–76. Amsterdam: Elsevier

OECD. 2006. *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006*. Paris: Org. Econ. Coop. Dev.

Pascarella ET, Terenzini PT. 2005. *How College Affects Students: A Third Decade of Research*. San Francisco: Jossey-Bass

Perie M, Moran R, Lutkus AD. 2005. NAEP 2004 trends in academic progress: three decades of student performance in reading and mathematics. *Rep. NCES 2005–464*, Washington, DC, U.S. Dep. Educ., Inst. Educ. Sci., Natl. Cent. Educ. Stat.

Porter A. 1995. The uses and misuses of opportunity-to-learn standards. *Educ. Res.* 24:21–27

Porter AC. 1993. School delivery standards. *Educ. Res.* 22:24–30

Rathbun A, West J, Hausken EG. 2004. *From Kindergarten Through Third Grade: Children's Beginning School Experiences*. Washington, DC: U.S. Dep. Educ. Natl. Cent. Educ. Stat.

Rosner J. 2003. On white preferences. *The Nation* 276(14):24

Rubin D, Stuart EA, Zanutto EL. 2004. A potential outcomes view of value-added assessment in education. *J. Educ. Behav. Stat.* 29:103–16

Sacks P. 2007. *Tearing Down the Gates: Confronting the Class Divide in American Education*. Berkeley: Univ. Calif. Press

Schmidt AE, Camara WJ. 2004. Group differences in standardized test scores and other educational indicators. See Zwick 2004b, pp. 189–201

Secr. Educ. Comm. Future High. Educ. 2006. *A Test of Leadership: Charting the Future of U.S. Higher Education*. Washington, DC: U.S. Dep. Educ.

Shepard LA. 1993. Evaluating test validity. *Rev. Res. Educ.* 19:405–50

Snyder TD, Dillow SA, Hoffman CM. 2006. *Digest of Education Statistics 2006*. Washington, DC: U.S. Dep. Educ., Inst. Educ. Sci., Natl. Cent. Educ. Stat.

Soares JA. 2007. *The Power of Privilege: Yale and America's Elite Colleges*. Stanford, CA: Stanford Univ. Press

Sorensen AB. 1989. Schools and the distribution of educational opportunities. *Res. Sociol. Educ. Socializat.* 8:3–26

Sorensen AB, Hallinan M. 1977. A reconceptualization of school effects. *Sociol. Educ.* 50:273–89

Sorokin PA. 1959. *Social and Cultural Mobility*. Glencoe: Free Press

Spearman C. 1904. "General intelligence," objectively determined and measured. *Am. J. Psychol.* 15:201–92

Sternberg RJ, Grigorenko EL, Kidd KK. 2005. Intelligence, race, and genetics. *Am. Psychol.* 60:46–59

Studley R. 2006. Inequality, student achievement and college admissions. See Zwick 2004b, pp. 321–43

Traiman SL. 1993. *The Debate on Opportunity to Learn Standards*. Washington, DC: Natl. Gov. Assoc.

U.S. Congr. Off. Technol. Assess. 1992. *Testing in American Schools: Asking the Right Questions. Rep. OTA-SET-519*. Washington, DC: USGPO

Warren JR, Grodsky E, Lee JC. 2008. State high school exit examinations and post-secondary labor market outcomes. *Sociol. Educ.* 81:77–107

Warren JR, Kulick R. 2007. Modeling states' enactment of high school exit examination policies. *Soc. Forces* 86:215–30

Weber M. 1978. Bureaucracy and education. In *Economy and Society: An Outline of Interpretive Sociology*, ed. G Roth, C Wittich, pp. 998–1001. Berkeley: Univ. Calif. Press

Young JW. 2004. Differential validity and prediction: race and sex difference in college admissions testing. See Zwick 2004b, pp. 289–301

Zwick R. 2004a. Is the SAT a "wealth test?" The link between educational achievement and socioeconomic status. See Zwick 2004b, pp. 203–16

Zwick R, ed. 2004b. *Rethinking the SAT: The Future of Standardized Testing in University Admissions*. New York: RoutledgeFalmer

Annual Review
of Sociology

Volume 34, 2008

Contents

Prefatory Chapters

Reproductive Biology, Technology, and Gender Inequality: An Autobiographical Essay
Joan N. Huber 1

From Mead to a Structural Symbolic Interactionism and Beyond
Sheldon Stryker 15

Theory and Methods

Methodological Memes and Mores: Toward a Sociology of Social Research
Erin Leahey 33

Social Processes

After Secularization?
Philip S. Gorski and Ateş Altınordu 55

Institutions and Culture

Religion and Science: Beyond the Epistemological Conflict Narrative
John H. Evans and Michael S. Evans 87

Black/White Differences in School Performance: The Oppositional Culture Explanation
Douglas B. Downey 107

Formal Organizations

Sieve, Incubator, Temple, Hub: Empirical and Theoretical Advances in the Sociology of Higher Education
Mitchell L. Stevens, Elizabeth A. Armstrong, and Richard Arum 127

Political and Economic Sociology

Citizenship and Immigration: Multiculturalism, Assimilation, and Challenges to the Nation-State
Irene Bloemraad, Anna Korteweg, and Gökçe Yurdakul 153

Differentiation and Stratification

The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets
Devah Pager and Hana Shepherd 181

The Second Generation in Western Europe: Education, Unemployment, and Occupational Attainment
Anthony F. Heath, Catherine Rothon, and Elina Kilpi 211

Broken Down by Race and Gender? Sociological Explanations of New Sources of Earnings Inequality
Kevin T. Leicht 237

Family Structure and the Reproduction of Inequalities
Sara McLanahan and Christine Percheski 257

Unconscious Racism: A Concept in Pursuit of a Measure
Hart Blanton and James Jaccard 277

Individual and Society

Horizontal Stratification in Postsecondary Education: Forms, Explanations, and Implications
Theodore P. Gerber and Sin Yi Cheung 299

Gender Inequalities in Education
Claudia Buchmann, Thomas A. DiPrete, and Anne McDaniel 319

Access to Civil Justice and Race, Class, and Gender Inequality
Rebecca L. Sandefur 339

How the Outside Gets In: Modeling Conversational Permeation
David R. Gibson 359

Testing and Social Stratification in American Education
Eric Grodsky, John Robert Warren, and Erika Felts 385

Policy

Social Networks and Health
Kirsten P. Smith and Nicholas A. Christakis 405

Sociology and World Regions

Gender in African Population Research: The Fertility/Reproductive Health Example
F. Nii-Amoo Dodoo and Ashley E. Frost 431

Regional Institutions and Social Development in Southern Africa
Matthew McKeever 453

Conditional Cash Transfers as Social Policy in Latin America: An Assessment of their Contributions and Limitations [Translation]
Enrique Valencia Lomelí 475

Las Transferencias Monetarias Condicionadas como Política Social en América Latina. Un Balance: Aportes, Límites y Debates [Original, available online at http://www.annualreviews.org/go/EValenciaLomeli]
Enrique Valencia Lomelí 499

Indexes

Cumulative Index of Contributing Authors, Volumes 25–34 525

Cumulative Index of Chapter Titles, Volumes 25–34 529

Errata

An online log of corrections to *Annual Review of Sociology* articles may be found at http://soc.annualreviews.org/errata.shtml