June 1982

Employment Tests and Employment Discrimination: A Dissenting Psychological Opinion

Craig Haney

Follow this and additional works at: https://scholarship.law.berkeley.edu/bjell

Recommended Citation

Link to publisher version (DOI)
https://doi.org/10.15779/Z38QP7R

This Article is brought to you for free and open access by the Law Journals and Related Materials at Berkeley Law Scholarship Repository. It has been accepted for inclusion in Berkeley Journal of Employment & Labor Law by an authorized administrator of Berkeley Law Scholarship Repository. For more information, please contact jcera@law.berkeley.edu.
Employment Tests and Employment Discrimination: A Dissenting Psychological Opinion

Craig Haney†

Employment tests are widely used in our society. The author critically examines the assumptions upon which such tests are based and analyzes related judicial decisions. He concludes that rigorous legal regulation of the test industry is needed and includes a series of policy recommendations designed to enhance the fairness of employment test practices.

Employment tests are creatures of a meritocracy. They represent a technology of individual measurement whose apparent purpose is the allocation of economic opportunity and rewards. This technology resonates perfectly with the individualism of our law, which has progressed almost completely "from status to contract," and supposedly operationalizes the abstract principle of political economy, "to each according to his ability." Many psychologists regard ability and

† Associate Professor of Psychology, University of California, Santa Cruz. B.A., University of Pennsylvania, 1969; Ph.D. in Psychology, Stanford University, 1978; J.D., Stanford University, 1978.

1. This technology is a major component in the American meritocratic vision. Consider Goslin:

In the United States concern with the measurement of abilities, and in particular, intellectual skills, has grown steadily during the past forty or fifty years. . . . In no other society do we find as great a preoccupation with accurate assessment of ability combined with the allocation of opportunity for advancement in the society on the basis of ability. . . . And we have certainly done more about developing systematic techniques for assessing abilities than any other society.


3. Consider Herbert Spencer, The Man vs. the State; A Collection of Essays (1884): "[E]ach adult gets benefits in proportion to merit, reward in proportion to desert, merit and desert being understood as ability to fulfill all the requirements of life." Id. at 67. For a philosophical argument that society does have a moral right to impose a rule of hiring on corpora-
employment tests as their profession’s finest and most significant contribution to social and economic progress. Yet these tests represent a most formidable barrier to equal opportunity and racial justice in the workplace.

In recent years, testing has been afforded a major role in the social and legal debate over racial equality. The issue of fair employment testing brings the ideologies of equality and meritocracy into apparent conflict, and forces adversaries to refine their commitments to each. It also puts the courts in direct contact with the technical area of psychological test theory, often leaving them at the mercy of test specialists or experts. Employment testing has become a major industry in this country, with interests of its own to promote in this debate. In the context of these competing ideologies and interests, I believe that examination of both the nature and consequences of employment testing will make clearer its regulatory mandate.

My intention is to examine this topic from a somewhat broader perspective than separate legal and psychological analyses of employment testing ordinarily do. But I wish to draw some narrower technical implications from this broad-based discussion as well. In this article, after briefly reviewing the historical development of employment testing, I examine the legal standards with which these tests have been evaluated and controlled. In particular, in Part II, I review the stan-


tions, and that “hiring by competence” is the rule we should use because it is “as just as human nature allows,” see Goldman, Justice and Hiring by Competence, 14 AM. PHILO. Q. 17, 28 (1977). Goldman’s discussion, however, assumes “the ability to judge competence according to qualifications,” id. at 22, taking as “given” most of what is in dispute in cases of discriminatory employment testing. See also R. Nozick, Anarchy, State, and Utopia (1974); J. Rawls, A Theory of Justice 75-90 (1971). See R. Wasserstrom, Philosophy and Social Issues 68-77 (1979), for a critical and incisive discussion of the notion that “ability” can be equated with “merit” and serve as a decisive claim to opportunity.

4. One author suggests that “it is probably safe to say that there are more ability tests being given annually in the United States than there are people.” D. Goslin, The Search for Ability: Standardized Testing in Social Perspective 54 (1963). There are several thousand psychological tests in print. See, e.g., O. Burus, Tests in Print II: An Index to Tests, Test Reviews, and the Literature on Specific Tests (1974); K. Chun, S. Cobb & J. French, Measures for Psychological Assessment (1975). Syndicated business columnist Sylvia Porter reports that the use of tests in industry had declined somewhat in recent years because of the fear of litigation, but that the tests are now making a “spectacular comeback.” Indeed, she suggests that “every major employer in the U.S. is at least experimenting again (or for the first time) with psychological tests.” Porter, Big Comeback for Psychological Tests, San Francisco Chron., May 9, 1979, at 28, col. 1. By one estimate, there are some 500 commercial testing organizations in the United States. Committee on Ability Testing, Ability Testing: Uses, Consequences, Controversies, Part 1, at 10 (1982). For the period 1978-1979, the Educational Testing Service reported revenues in excess of $94 million. Fiske, Student Testing Unit’s Expansion Leads to Debate, N.Y. Times, Nov. 14, 1979, at A1, col. 2. Harcourt, Brace, & Jovanovich, which acquired The Psychological Corporation in 1970, reported their 1979 revenues in “tests and testing services” at almost $30 million. See Sokal, The Origins of The Psychological Corporation, 17 J. HIST. BEHAV. SCI. 54 (1981). But see Tenopyr, The Realities of Employment Testing, 36 AM. PSYCHOLOGIST 1120 (1981), who suggests that the importance of employment tests is “probably overrated” since testing is “not universal” in either the public or private sectors. Id. at 1120.
standards that the Supreme Court has applied to testing and discuss the possibility that these standards may be in the process of radical but ill-advised change. Part III contains a discussion of some basic testing concepts and a critical evaluation of their current application in employment discrimination law. The logic and utility of employment testing is affected by broader changes in psychological theory and social conditions, and several of these changes are examined in Part IV. Finally, in Part V, I make a series of policy recommendations designed to enhance the fairness of employment test practices.

I. TESTING IN CONTEXT: A BRIEF HISTORY OF THE DEVELOPMENT OF EMPLOYMENT TESTS

The contemporary functions and limitations of employment testing are best clarified if first placed in historical context. A brief analysis of the conditions under which they were “discovered” and developed will provide some insight into their original purpose and the reasons for their continued use.5

Contrary to testing’s current popular image, the history of employment testing is not the history of a neutral and apolitical scientific method. Rather, the technology of testing originated in a theory of human behavior that was politically charged and laden with ideology. Critical stages in the development of testing were influenced by motives and conditions that compromised objectivity and careful evaluation. To be sure, this history should not be decisive in contemporary debates over the place of testing in a policy of fair employment. Neither should it be forgotten.

Although “employment tests”6 originated with the practices of the Chinese civil service over 3,000 years ago, modern testing has its roots in the psychological theories of the late nineteenth century. Individualism was the prevailing ethos of that time, and the measurement of individual differences preoccupied the emerging human and psychological sciences in both Europe and America. The first and most influential psychometricians—Galton, Pearson, and later, R.A. Fisher—were eugenicists. Their work originated in the Eugenics Laboratory of University College, London, and concentrated on issues of evolution and

---


6. The term “employment test” is used throughout this article to refer to any test whose results are used in making any employment decision. These tests are thus distinguished from other psychological tests by their use, rather than by the characteristics of the tests themselves.
biological inheritance. They were strong believers in the importance of genetic makeup as a determinant of human traits and abilities, and concerned themselves with the measurement of all forms of individual differences as a method of calculating the effect of heredity on behavior. Their techniques were transported to America by James Cattell, who began the Psychological Laboratory at Columbia in 1891.

The other major figure in the early development of test methods was Alfred Binet. A French lawyer who turned to the study of personality and psychopathology, Binet became interested in the systematic assessment of individual differences, especially cognitive abilities. He devised the first “intelligence test” as a method of diagnosing mental retardation in children who were to be sent to special institutions rather than remain in the Paris school system. The original purpose of diagnosing retardation was soon extended to include the categorization of all students into ability groupings. Binet’s test achieved widespread use not only in Europe but in the school systems of the United States as well. Several years after its initial success in 1905, Binet proposed that his testing method be used to aid in determining the responsibility of criminals and in classifying inductees into the Army.

Binet was unusual among the early test developers in that he took no stand on the causes of performance differences in “intelligence” test scores. He certainly did not believe that intelligence was inherited in fixed quantities, unmodifiable through training, and he termed such a position “brutal pessimism.” However, this brutally pessimistic view was widely adopted by American psychologists. Binet’s test was quickly adapted in this country by Lewis Terman and others for use in mea-


The legacy of the early English geneticists can be seen in Great Britain’s long-standing preoccupation with intelligence testing and emphasis on the contribution of heredity to performance. Cyril Burt is an extreme but well-known example: “[I]mproved environmental amenities can of themselves ensure no lasting results; but the changes in a nation’s genetic constitution are likely to prove irreversible.” Burt, The Inheritance of Mental Ability, 13 AM. PSYCHOLOGIST 1, 9 (1958). See D. Goslin, supra note 4, at 39. There is an American legacy as well. Compare, e.g., Garth, A Review of Racial Psychology, 22 PSYCHOLOGICAL BULL. 353, 359 (1925), with R. Herrnstein, I.Q. and the Meritocracy (1973), and A. Jensen, Genetics and Education (1972).

8. See, e.g., Cattell & Farrand, Physical and Mental Measurements of the Students of Columbia University, 3 PSYCHOLOGY REV. 618 (1896). There was nearly universal acceptance of eugenics by early American psychologists of all political persuasions. Cattell, who founded The Psychological Corporation in 1921, conceived of himself as a “radical democrat,” and corresponded with Samuel Gompers about how psychology could contribute to unionism. Sokal, supra note 4, at 58.

suring “innate intelligence” and detecting “genetic inferiority.” As for the latter, Terman predicted that “in the future intelligence tests will bring tens of thousands of these high grade defectives under the surveillance and protection of society.”

The enthusiasm for individual measurement led to the development of tests for traits and abilities other than intelligence. The United States Civil Service Commission began a program of measuring the skills and aptitudes required for specific jobs as early as the 1880’s. But the first real use of ability tests in American industry began with the work of Hugo Münsterberg. Münsterberg was among the first to realize the commercial applications of the measurement techniques that were being developed in academic psychological laboratories. In 1910, he surveyed several hundred business executives to elicit from them a list of the psychological traits they believed to be necessary and desirable in employees. Münsterberg built on the techniques of measurement that Binet had devised and was soon administering selection tests in industry, helping the American Tobacco Company select its traveling salesmen and the Boston Elevated Company its conductors.

Business management during these times, concerned that inefficient hiring techniques were reducing profits, turned increasingly to “science” for assistance. Eventually Münsterberg’s ideas were taken seriously by the academic community as well, and in 1915 the Carnegie Institute of Technology began a Division of Applied Psychology that performed psychological consultations for industry on employment selection.

10. L. Terman, The Measurement of Intelligence 6 (1916). He also commented on the low scores achieved by minorities:

Their dullness seems to be racial, or at least inherent in the family stocks from which they come . . . Children of this group should be segregated in special classes . . . There is no possibility at present of convincing society that they should not be allowed to reproduce, although from a eugenic point of view they constitute a grave problem because of their unusually prolific breeding.

Id. at 91-92. “Scientific” racism was by no means restricted to psychologists. See, e.g., J. Haller, Outcasts from Evolution: Scientific Attitudes of Racial Inferiority, 1859-1900 (1971).


12. Münsterberg is in many respects the founder of “law and psychology.” See, e.g., H. Münsterberg, On the Witness Stand (1908) (using psychological test techniques and concepts of memory and association for examination of witnesses).


14. Münsterberg urged psychologists to take their techniques into the “real world.” However, he was also a staunch believer in the “detached and dispassionate” brand of applied psychology. He argued that psychologists should tell industry what psychological effect a particular course of action would have, but that they had “no right to decide which effect is good and which effect is bad.” H. Münsterberg, Business Psychology 182 (1915). On Münsterberg’s remark-
The success of employment testing was linked to the ability of its proponents to convince industry of its cost effectiveness. Indeed, the fate of "industrial psychology" itself was linked to this ability, since this applied discipline had begun with the use of psychological tests in industry and the tests continued to be its most marketable product:

In contrast to the European acceptance of industrial psychology, the American experience was confused and disorganized. . . Industrial psychology had to make its own way. With little aid, industrial psychologists had to persuade management of their utility. A few hardy men persisted in their effort to market their knowledge. The most tangible instrument they had developed was the psychological test, and on the success of this tool industrial psychologists were to stake their future.15

Almost from its inception, then, the employment test was a business product as well as a scientific technique. As with all products, its success depended as much on what buyers thought it could do as on what it could, in fact, do.

The major impetus for the systematic development of ability and aptitude testing, however, came from the First World War and government subsidization of psychologists to develop army classification programs. "[T]he modern era of personnel selection had its beginnings on April 6, 1917, when the United States declared war on Germany."16 Robert Yerkes was president of the American Psychological Association when the war began, and he organized a committee of psychologists to determine what contribution the profession might make to the war effort.17 The committee decided that psychological tests offered the best possibility for practical service. Large numbers of people had to be assessed, quickly and inexpensively. To meet these needs, a "group" test of intelligence was created whose questions were standardized and presented in written form. So efficient were the techniques that more than 1.7 million men were tested and classified in less than two years.

When the war ended, many of the techniques that had been developable career, see M. Hale, Human Science and Social Order: Hugo Münsterberg and the Origins of Applied Psychology (1980); Moskowitz, Hugo Münsterberg: A Study in the History of Applied Psychology, 32 Am. Psychologist 824 (1977). Münsterberg had no doubts about the genetic and unmodifiable nature of the abilities his tests were designed to measure. "We can learn to keep house with our resources," he wrote, but even scientific attempts to improve on those resources "cannot deceive us as to the fundamental fact that the decisive tendencies of our mind are inherited and cannot be fundamentally changed." Quoted in M. Hale, supra, at 124.

15. L. Baritz, supra note 5, at 57.
17. Yerkes had been Münsterberg's student and assistant, and had prior commitments to testing and the genetic view of behavior it then represented. He had developed a "point scale" version of the Binet "I.Q." test, served on the Committee on Eugenics of the National Commission on Prisons, and in 1917 became the chairman of the Eugenics Research Association's Committee on Inheritance of Mental Traits.
EMPLOYMENT TESTS AND DISCRIMINATION

opposed were applied directly to peacetime use. Indeed, the war proved a boon to the fledgling test industry. The Army's group testing "convincing the nation that adequate prediction of success could be achieved through mass processing, and schools and industry were quick to demand tests of this type after the war." What was interesting about the wartime achievement was that its actual effects were not very carefully assessed. Aside from the speed and economy of the testing itself, and the obvious fact that classification decisions of some sort could be based on the results, the practical value of the program was not calculated. As one later enthusiast conceded about the tests: "The point that is important to note is the blind faith which many people displayed in the results obtained from them." While the tests allowed a massive program of classification to be implemented quickly and cheaply, the accuracy and utility of the classifications were more assumed than demonstrated.

Nonetheless, many of the wartime psychologists were highly sought after as industrial consultants, and pursued subsequent careers in testing. Moreover, the testing experience affected the profession of psychology generally. "The greatest impact of the Army program was on psychologists and their concept of the role of psychology in society. Before World War I psychology was largely an academic discipline; thereafter it became more and more a profession."

Psychology was transformed and invigorated by the wartime testing program. One historian observed that "the advertising that this testing gave psychology in America reached into the remotest corner of the laboratory and swelled college classes, creating a great demand for Ph.D. instruc-

18. L. Cronbach, Essentials of Psychological Testing 162 (2d ed. 1960). See also Sokal, supra note 4, at 61: "Mental testing emerged as a social movement after the apparent success of the Army testing program, and some psychologists set out to try to influence all aspects of American life." (Footnotes omitted.) But the test results were not accepted uncritically by everyone. See, e.g., Pastore, The Army Intelligence Tests and Walter Lippmann, 14 J. Hist. Behav. Sci. 316 (1978).

19. The "evidence" marshalled in support of the value of the army testing program was often no more sophisticated than this:

Our army abroad had a well earned reputation for efficiency and no small part of the result may be attributed to the fact that the lowest 10 per cent in intelligence were not sent overseas and that 83 per cent of the officers came from the "A" and "B" classes—superior and very superior intelligence.

H. Goddard, Human Efficiency and Levels of Intelligence 34 (1920).


21. P. Dubois, supra note 5, at 67. "The conspicuous success of the program engendered confidence in measuring new variables and applying the results not only in schools and child guidance clinics but also in vocational counseling and in the selection of industrial personnel."

Id.
The fate of the profession was in many ways tied to the success with which test specialists marketed their tool for peacetime use.

One of the important peacetime uses of the tests was in the movement to restrict immigration into the United States. Intelligence tests administered at the Ellis Island receiving station in New York in 1912 had already "documented" the fact that fully four-fifths or more of the Jews, Hungarians, Italians, and Russians entering this country were "feeble-minded."\(^2\) After the war, however, the tests' new-found status gave psychologists a more influential role in immigration policy-making. The major "scholarly" contribution to this movement was written by Carl Brigham (who went on to develop the Scholastic Aptitude Test) and was based on the Army data collected by psychologists in their wartime testing.\(^2\) In the foreword to the book, Robert Yerkes suggested that "no one of us as a citizen can afford to ignore the menace of race deterioration or the evident relations of immigration to national progress and welfare."\(^2\) Brigham himself warned of a decline in the quality of immigrants who were entering the country and attributed their increasingly poor test scores to a decrease in the proportion of Nordic blood that the recent immigrants possessed.\(^2\)

Although most of the discussions of the Army test data focused on the performance of immigrant groups, results for American blacks were

\(^{22}\) E. Boring, A History of Experimental Psychology 575 (2d ed. 1950).
\(^{24}\) C. Brigham, A Study of American Intelligence (1923).
\(^{25}\) Yerkes, Foreword to C. Brigham, supra note 24, at vii.
\(^{26}\) C. Brigham, supra note 24, at 171. Brigham based these assertions on his "discovery" that test performance was a direct function of time spent in the United States (i.e. those who had immigrated 20 years before performed better as a group than those who had immigrated more recently). He rejected the obvious implication that this finding had for the cultural bias of the test itself, interpreting it instead as evidence that the more recent immigrants were genetically inferior to the earlier ones. He also wrote:

If the tests used included some mysterious type of situation that was "typically American," we are indeed fortunate, for this is America, and the purpose of our inquiry is that of obtaining a measure of the character of our immigration. Inability to respond to a "typically American" situation is obviously an undesirable trait.


\(^{27}\) See L. Kamin, supra note 5, for an account of the role played by psychological testing in the passage of this legislation. See also Samelson, On the Science and Politics of the I.Q., 42 Soc. Research 467 (1975), and Kamin's Reply to Samelson, 42 Soc. Research 488 (1975).
presented as well. This was the first extensive indication that blacks performed poorly on these tests. Interestingly, however, this poor performance was little discussed. In contrast to the shock with which the immigrants' low scores were greeted, the blacks' performance seemed to be expected and, therefore, in need of little explanation. It was the data of the white native draftees, for example, that was routinely separated out and used in the comparisons with the scores of foreign born draftees. Brigham's analysis used the scores of the black draftees as the lowest baseline or yardstick against which the lack of immigrant intelligence was assessed. Brigham concluded that race deterioration was a far greater problem in America "for we are incorporating the negro into our racial stock, while all of Europe is comparatively free of this taint."

These racist implications were perfectly consistent with the eugenic origins of the tests. The war had simply increased the data base from which racial and ethnic comparisons could be made, and enhanced the prestige of the psychologists who made them. Testing was used as the instrument of a racist world view that held whole groups of people to be genetically inferior to others, while the early test enthusiasts proclaimed the neutrality of the instruments that supposedly documented racial inferiority. These proclamations were premised in part on a genuine naiveté about the tests themselves, and in part on the fact that the tests told proponents what they either "knew" or very much wanted to hear. Whether test developers and administrators created and employed these tests with the express purpose or intention of discriminating against minorities is largely beside the point—tests were used to put a scientific and documentary gloss on racist social policies. Test psychologists most certainly did not originate the racist sentiments that found expression in harsh and discriminatory immigration laws. Nor were they the most extreme purveyors of these doctrines. But they were the only ones who could employ a "scientific" instrument to support their claims.

28. R. YERKES, PSYCHOLOGICAL EXAMINING IN THE UNITED STATES ARMY 693-99 (1921).

29. C. BRIGHAM, supra note 24, at 144.

30. Id. at 209. In this context, consider Samelson, supra note 20, who reports that the Army policy of giving an individual examination to any recruit who failed the group test was not followed for the majority of blacks who failed the "Beta test" (the test given to those who could not read or understand English): "Presumably under time pressure and lacking in manpower, the testers thought it unnecessary, since the outcome was obvious anyway—even though there is some evidence that when individual examinations were given to Negroes, the results showed a noticeable improvement . . . ." Id. at 279-80. Moreover, there was some evidence that blacks who could understand but not read English were so confused and mystified by the completely nonverbal pantomime instructions used to administer the Beta test "that they dozed off en masse." Id. at 280 (footnote omitted). Yet, as Samelson notes, "the summary reports, often repeated later, gave only the cold scientific fact that . . . . 'the intellectual status of the Negro [was] greatly inferior to that of the white.'" Id. (footnote omitted).
The next several decades were characterized by a proliferation of many new tests to measure new and different individual traits and abilities. The Depression years curtailed somewhat the extent to which American employers used tests for personnel screening, but psychologists were not deterred from inventing new ones. During this period, however, there were virtually no important breakthroughs or changes in the fundamental logic or method of psychological testing. In the preface to a 1942 edition of a widely used text in employment selection, for example, its author wrote: "There has been much activity in the field of personnel psychology since the first edition of this book (in 1926). A little of this activity has involved the development of new principles such as factor analysis. Most of the work, however, has consisted of more extensive use of existing principles." 

After a period of comparatively little interest in or subsidization of psychological testing methods, the government poured large amounts of money into the development of testing programs for military use in World War II. Again, the major impetus for the widespread and systematic development of tests came from the need to process and classify huge numbers of people in a military draft. The emphasis was on expediency: "All the services wanted measures—quick ones—of abilities and of aptitudes." Large numbers of psychologists devoted much energy to the development of tests under wartime conditions. This time, many more persons were tested and many more psychologists were involved. As one of them put it, "World War II can . . . be said to have been a social scientist's war." The legitimacy of psychological tests was enhanced, along with the prestige of the social scientists who developed and administered them. "Psychology in America came into its own in the Second World War, and it was largely, at least at first, by way of the testing."

There can be no question that testing had "caught on." The Psychological Corporation, the profession's largest test distributor, noted that "an adequate number of professionally trained persons simply is

31. One review found that over 5005 testing articles were published in the fifteen-year period between 1921 and 1936. Most of them were reports of new tests. E. South, An Index of Periodical Literature on Testing, 1921-1936, at iii (1937). For two glimpses of psychology in the Depression years, see Finison, Unemployment, Politics, and the History of Organized Psychology, 31 Am. Psychologist 747 (1976); Risse, Vocational Guidance During the Depression. Phrenology Versus Applied Psychology, 12 J. Hist. Behav. Sci. 130 (1976).
33. E. Boring, supra note 22, at 577.
34. It was estimated that in 1944 approximately 60 million standardized tests were given to approximately 20 million Americans. Of these totals nearly half were accounted for directly by the military. Wolfe, Testing is Big Business, 2 Am. Psychologist 26, 26 (1947).
35. Darley, Five Years of Social Science Research, in Groups, Leadership, and Men 3, 12 (H. Guetzkow ed. 1951).
36. See L. Baritz, supra note 5, at 139-66.
37. E. Boring, supra note 22, at 577.
not available to fill all the legitimate psychological positions now open." In 1947, the American Psychologist proclaimed that testing was "big business," and few could argue. By one account well over half of the American psychologists in 1948 were engaged in testing work of some kind. The expediency and convenience of the tests were attractive to post-war employers faced with a large supply of workers whose recent experience was of uncertain utility. The use of such tests by industrial psychologists, as well as the use of industrial psychologists by American industry, was solidified. Yet "the single greatest pitfall was the unchecked enthusiasm and naivete with which American industrial managers accepted the principles and practice of psychological testing."41

The enormous growth of the test industry did not result from any sudden increase in sophistication on the part of test developers or their techniques of measurement. In fact, the war had produced no major advances in basic test methodology over the group techniques that had been invented in the earlier war. The effort "of the groups who carried on the psychological work of World War II is clearly evolutionary, involving no great changes in testing procedures and no wholly new statistical methods."42 Indeed, testing has continued to grow in use and application, but this growth has not been the product of any significant technical breakthroughs or radical changes in method: "What improvements have been made have resulted from refinements of existing

38. Bennett & Seashore, Checking the Qualifications of Purchasers of Tests, 1 AM. PSYCHOLOGIST 353, 356 (1946).
39. Wolfe, supra note 34, at 26. The "big business" aspect was recognized by some members of the profession as a distinctly mixed blessing. In 1950, members of an American Psychological Association [hereinafter APA] committee noted that "in the case of any extensive business operation in which the actual and potential profits are great, ethical problems arise and are at times difficult to resolve." APA Committee on Ethical Standards for Psychology, Ethical Standards for the Distribution of Psychological Tests and Diagnostic Aids, 5 AM. PSYCHOLOGIST 620, 620 (1950). Among the problems they identified were (1) "the overselling of tests" by test industry representatives "who are virtually untrained in testing," id. at 624; (2) "publication of a test by an organization which lacks orientation to the technical problems of test construction and use, and which does not have the type of sales organization which can most ethically market tests," id.; (3) publication of tests "motivated by the desire for the recognition, the prestige, and the income which result from the publication of tests even when inadequately standardized," id. at 625; and (4) the biased selection of data included in test manuals and publications by psychologists who wish their tests to appear "in the best possible light," including incidents in which "the lack of data may be glossed over and disguised," as well as "the over-simplification of data concerning reliability and validity, in an effort to make the test appeal to unsophisticated users and to broaden its possible application," id.
40. F. Goodenough, supra note 20, at 90.
41. L. Baritz, supra note 5, at 156. Baritz suggests that "[a]s if to compensate for decades of apathy and skepticism, managers now overreacted, as they had in the post World War I period." Id. Fortune commented, for example, that "[m]ost of the errors in industrial psychology... have been committed by test-happy managements that have not taken the trouble to check up on the tests they use." The Tests of Management, FORTUNE, July 1950, at 92, 92.
42. F. Goodenough, supra note 20, at 502.
techniques or methods or from the abandonment of unreliable, invalid techniques and instruments . . . rather than from any new approaches or philosophies.\textsuperscript{43} Extensive resources have been devoted to test development and marketing and the methodology is employed on a more widespread basis than ever before. Yet, as one historian of psychological testing wrote:

Testing by psychologists began in the nineteenth century as the work of relatively isolated individuals. . . . Much of the energy of recent decades has gone into modest improvements in existing measures and relatively slight modifications in their applications.\textsuperscript{44}

Thus, the intellectual origins of modern testing lay in eugenics and a belief in the importance of innate characteristics as determinants of behavior and performance. The ideology of racial superiority and domination was part of the testing movement from its inception. At the same time, the application of testing to industry was accomplished in the name of profit and in the belief that "scientific" selection could conveniently reduce the costs of hiring errors. The major impetus and opportunity for scientific development, however, came in wartime, when psychologists were enlisted in the task of categorizing large numbers of inductees quickly and cheaply, and under conditions in which precipitous rejection of people for military jobs was tolerable because of a huge labor supply.

From one perspective, then, the history of testing is the history of an ideology, a business product, and a military expedient. This is a side of the test industry that is not often presented. Modern employment and ability tests have acquired all of the trappings of a highly developed technology of individual measurement, premised on scientifically neutral principles. Yet the fundamental logic of the enterprise has not changed much since Galton’s time. The first testers were fascinated with the idea that individual ability, merit, and worth could be measured. However, they were far less diligent in examining defects in their tests than alleging defects in the people who took them. Proponents of modern testing certainly do not all share the ideology of ge-

\textsuperscript{43} Wallace, \textit{Criteria for What?}, 20 Am. Psychologist 411, 412 (1965). He goes on to suggest that:

the hope that these techniques would lead to equally accurate predictions in the less well-structured, controlled, and specific situations of post-training performance in the real world of business and industry has not been fulfilled. The predictive test validities reported for vocations today are little, if any, higher than those reported 15 years ago.\textsuperscript{Id}

\textsuperscript{44} P. DuBois, supra note 5, at 131. Consider, finally, the observations of Oscar Buros, looking back on a long and distinguished career in psychological testing: "We don't have a great deal to show for fifty years of work . . . The improvements—except for the revolutionary scoring machines and computers—have not been of enough consequence to permit us to have pride in what we have accomplished. . . . In fact, some of today's tests may even be poorer [than those of fifty years ago]." Buros, \textit{Fifty Years in Testing: Some Reminiscences, Criticisms, and Suggestions}, 6 Educ. Researcher 10 (1977).
netic determinism so common among their predecessors. They are far more sophisticated about the limits of the testing methods and techniques they use. But the legal community would do well to attend to the history of testing, its early motives, and the conditions under which it developed. It is a history that should not be discounted when the neutrality and scientific precision of these devices is proclaimed.

II. Developing Legal Standards: The Testing Trilogy

The early years of employment testing were notable for the lack of discussion about test discrimination and the differential effects that test instruments had on minority groups. Unquestioned faith in the accuracy of these instruments, the usually unstated belief on the part of the test administrators in the existence of racial differences, and a relative lack of concern on the part of most employers about the disparate racial impact of their hiring and employment procedures all contributed to the failure to examine racially discriminatory test practices. For employers who did have an intent to discriminate, differential test scores were “scientific proof” that they were justified. And for many of those who did not, tests still appeared to be the most neutral and objective selection devices one could use. Apparently unmindful of the discriminatory impact of testing itself, some test psychologists offered their product as a solution to racial injustice in the workplace. Indeed, compared to the egregious forms of blatant and overt discrimination then commonplace, tests seemed fairer.

Following Brown v. Board of Education, however, increasing attention was paid to the academic performance of minorities. Disparities in test performance were more widely publicized and their ultimate causes widely debated. There was a growing recognition that the source of racial disparities in test performance resided in the racist practices of the society at large. There came an increased awareness that testing might play a role in actually creating and increasing the very disparities that the tests appeared only to measure. This sentiment coincided with the publication of several widely discussed critiques of the test industry that challenged the propriety of psychological testing.

45. I do not mean to imply that professional psychology continued to be dominated by the racist worldview that prevailed during the late nineteenth and early twentieth centuries. Indeed, see Samelson, From “Race Psychology” to “Studies in Prejudice”: Some Observations on the Thematic Reversal in Social Psychology, 14 J. Hist. Behav. Sci. 265 (1978). Employment testing is one area, however, where racially disparate test results continued to be accepted with a relative lack of critical analysis reminiscent of the earlier period. See infra notes 133-34 and accompanying text.


47. 347 U.S. 483 (1954). A discussion of the role played by the social science evidence in Brown may be found in P. Rosen, The Supreme Court and Social Science 134-74 (1972).
generally and questioned the legitimacy and expertise of test administrators and developers as well. Careful legal scrutiny was to follow.

A. The Early Cases: Struggling with "Discriminatory Use"

Title VII of the Civil Rights Act of 1964 prohibited discriminatory employment practices. Although originally based on the assumption that employment discrimination was the result of relatively isolated cases of individual or organizational prejudice or ill-will, the law was intended to eliminate all unjustified impediments to equal employment in whatever form those impediments might take. Title VII did not forbid the use of employment tests, and said so explicitly in section 703(h):

[I]t shall not be an unlawful employment practice for an employer . . . to give and to act upon the results of any professionally developed ability test provided that such test, its administration or action upon the results is not designed, intended or used to discriminate . . . .

Test psychologists regarded this provision tentatively and with skepticism. Although lawmakers had assured a representative of the American Psychological Association that the provision "did not mean anything," other psychologists were justifiably concerned. One was worried that testing would be hampered by ensuing litigation. Another predicted that personnel managers would turn increasingly to tests as a pretext for discrimination. And one saw the provisions as an occasion for psychologists to remedy bad testing practices, including “the all-too-common practice of taking a convenient brief intelligence test off the shelf and using it for all jobs . . . frequently without any

48. E.g., H. Black, They Shall Not Pass (1963); M. Gross, The Brain Watchers (1962); B. Hoffmann, The Tyranny of Testing (1962). The general public criticism, however, focused more on personality tests than on measures of job skill or intellectual ability, and racially discriminatory aspects of the tests were of far less concern than basic privacy issues. Many persons objected to the intrusive quality of the tests and the personal nature of the information sought, while others expressed doubts that the information could be interpreted and used by psychologists in a meaningful way. An entire issue of the American Psychologist was devoted to a discussion of the Congressional hearings precipitated by this public outcry. See Amrine, The 1965 Congressional Inquiry into Testing: A Commentary, 20 Am. Psychologist 859 (1965).


52. Amrine, supra note 48, at 860.


demonstrable relevance to the selection problem at hand.”

The task of interpreting Title VII—including its testing provision—belonged to the courts and to the Equal Employment Opportunity Commission (EEOC). As the enforcement agency for Title VII, the EEOC in 1966 issued a brief, preliminary set of guidelines on employment testing, prepared with the assistance of a panel of psychologists outside the agency. These guidelines were followed by the Department of Labor’s standards, and a revised version of the Civil Service Commission’s guidelines for promotion. After a half century of virtual silence, federal agencies were struggling to formalize a testing policy sensitive to racial impact. Everyone agreed that discriminatory testing was illegal. There was, however, a wide diversity of opinion about what constituted discriminatory testing.

The lower federal courts offered conflicting interpretations of section 703(h). Disagreement centered on exactly what practices would amount to the “discriminatory use” of a professionally developed employment test. Some courts took the position that only tests that were discriminatorily administered or graded were barred by the Civil Rights Act. Others gave the section a more expansive reading by holding that any test that excluded a disproportionate number of minorities was prima facie discriminatory. These early cases also revealed important differences in the way courts understood the concept of “test validation,” as well as variations in how much importance they attached to it. Some courts felt that validation studies were unnecessary, while others indicated that they were essential in cases where minorities had been statistically underrepresented. In one case, the court’s own examination of the challenged tests led to the conclusion that the tests were “well within the capabilities of any person, regardless of his race, who has intelligence and alertness enough to work in

---

55. Id. at 802.
62. Test validation refers to a demonstration of the degree to which a test measures what it purports or is used to measure. See infra notes 172-86 and accompanying text.
the telephone industry."\textsuperscript{65}

One beneficial aspect of these cases, despite the confusing standards and relative naiveté concerning test methodology contained in the opinions, was the light they shed on actual testing practices and on their discriminatory impact. Until the testing practices of American business were scrutinized under Title VII, it was impossible to calculate their effects on minorities or, indeed, to even know what testing practices were in use. Goslin, a pre-Title VII testing researcher, describes the situation:

\begin{quote}
[I]t is difficult to get a clear picture of industrial testing because of the tremendous diversity of practice. Not only do companies as a whole vary with respect to testing, but frequently testing practices are different in the various branches, divisions, subsidiaries, and even departments . . . The result is that, in the absence of a stated company policy on testing, it is extremely difficult to evaluate the impact of standardized tests on personnel selection even where tests are widely administered.\textsuperscript{66}
\end{quote}

The early testing cases helped make public both the widespread use of employment tests and the enormous disparities in white and minority performance. Blacks' failure rates on some employment tests were three, four, and more times those of whites and, as a result, they were being denied jobs to the same extent.\textsuperscript{67}

\section*{B. Strict Scrutiny Under Title VII: Griggs and Albemarle}

The Supreme Court finally confronted the issue of employment testing under Title VII in \textit{Griggs v. Duke Power Co.}\textsuperscript{68} Its opinion resolved the testing controversy substantially in favor of the EEOC guidelines, lending considerable authority to the Commission's interpretation of Title VII's testing provision. The case was brought by thirteen black workers challenging, \textit{inter alia}, two employment tests that the company required for advancement into its more desirable departments, ones from which blacks had been openly excluded before passage of the Civil Rights Act. Approximately 58\% of the whites had passed these tests as compared to 6\% of the blacks. Moreover, the company had conducted no validation studies with which to evaluate the tests' job-relatedness. Noting that nothing in Title VII precluded the use of employment tests, the Court pointed out that "[w]hat Congress

\textsuperscript{66} D. Goslin, \textit{supra} note 4, at 95-96.
\textsuperscript{67} \textit{E.g.}, in \textit{Hicks}, the company employed two tests for persons seeking certain transfers into more desirable positions. The pass rates on these tests were 37.3% and 64.9% for whites, and 9.8% and 15.4% for blacks. 319 F. Supp. at 318. In \textit{Arrington v. Massachusetts Bay Transp. Auth.}, 306 F. Supp. 1355 (D. Mass. 1969), the company placed job applicants on a waiting list in order of their test scores. While 75\% of the whites scored in the range which gave them a reasonable chance of a job, only 20\% of the black test takers did.
\textsuperscript{68} 401 U.S. 424 (1971).
EMPLOYMENT TESTS AND DISCRIMINATION

has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance." Thus, the Court held that when tests excluded proportionately more blacks than whites, in an area whose current labor force still reflected a racial imbalance due to pre-Title VII discrimination, and when there was no relationship demonstrated between test scores and job performance, the tests were unlawful.

The decision was greeted with a storm of criticism. Foes predicted that it would result in the end of employment testing. They argued that it imposed unrealistic and impossible standards upon employers, ones that would force them to abandon testing programs altogether. Others claimed that the opinion was vague, and left too many unanswered questions.

In *Albemarle Paper Co. v. Moody*, the Court addressed many of these "unanswered" questions, but provided little solace to critics of *Griggs'" unreasonable" standards. *Albemarle* was brought by former and present black workers at a North Carolina paper mill. As in *Griggs*, at issue was the use of two employment tests to screen applicants into the company's better paying skilled positions. Under the "segregated lines of progression" policy that prevailed in the company before 1968, these positions had been available only to whites. Use of the employment tests now accomplished basically the same thing.

In *Albemarle* the Court reaffirmed a procedural requirement first articulated in *Griggs*: once the complainant has demonstrated the discriminatory impact of an employment testing procedure, the burden...
shifts to the employer to demonstrate that the tests are "job-related." Albemarl e’s main contribution, however, was to clarify standards of job-relatedness. Generally, the Court ruled, a test is job-related if it correlates significantly with important elements of the relevant work behavior. Moreover, job-relatedness must be demonstrated, and "cannot be proved through vague and unsubstantiated hearsay." The facts of Albemarle provided the Court with an opportunity to comment more specifically on various methods of demonstrating job-relatedness. The Court was highly critical of the "odd patchwork" of statistics taken from a study conducted on "the eve of the trial" by an industrial psychologist hired to evaluate the job-relatedness of the challenged testing program. Its critique articulated as yet unexamined aspects of employment test validation. First, the Court conceded that employers could use a test to screen applicants for a job other than the one for which the test was professionally validated, but only if there were "no significant differences" between the two jobs. Secondly, the Court encouraged the use of precise criteria of job performance in validation studies, rather than vague and subjective supervisorial ratings. Next, it suggested that performance criteria used in validation studies could be derived primarily from jobs near the top of a line of progression, rather than only at the entry level job. But this could be done only if advancement to the top was nearly automatic and the time of advancement was not so great as to permit significant changes in perform-


78. 422 U.S. at 428 n.23.
79. The district court had approved a validation study in which no job analysis had been done, the use of tests in six lines of progression for which no validation study had been done, and a requirement that two tests be passed to enter seven different lines of progression even though only one of those tests had been validated for those lines. Moody v. Albemarle Paper Co., 474 F. 2d 134, 138, 139 (4th Cir. 1973).
80. 422 U.S. at 432. The study compared current employees' test scores with judgments by their supervisors as to how they would perform in ten different job groupings. Id. at 411. The district court ruled that the tests had thus been validated and sustained the defendant employer's burden of proof. Id. Justice Stewart, writing for the Court, questioned the reliability of the individual tests, denying that the correlation of factors was statistically significant. Id. at 432.
81. Because there was no analysis of the job skills required for the positions referred to in the test, the Court concluded that the required finding of "no significant differences" could not have been made. Id.
82. Id at 432-33. The Court quoted the relevant EEOC Guidelines, 29 C.F.R. §§ 1607.5(b)(3)-(4), emphasizing the level of specificity required therein and comparing them unfavorably with the methods used by defendant Albemarle. Id. at 432-33 & n.30.
Finally, quoting testing standards of the American Psychological Association (APA) to the effect that a test should be validated on persons as similar as possible to those with whom it will be used, the Court suggested that differential validity studies be done on minority groups wherever technically feasible.

*Albemarle* went further than simply discussing standards for demonstrations of job-relatedness. The Court also noted that a demonstration of job-relatedness was not necessarily sufficient to end a Title VII inquiry. If the complaining party was able to show that alternative selection procedures would have a less discriminatory impact, the burden shifted back to the employer to prove that his tests were not being used merely as a "'pretext' for discrimination." Thus, a test having discriminatory impact might be technically job-related but still impermissible under Title VII if some other, less discriminatory, test could have been used by the employer.

Both *Griggs* and *Albemarle* appeared to establish clear trends in the Court's application of Title VII to employment tests that had a disparate impact on minorities. They represented increasingly explicit endorsements of the EEOC guidelines interpreting the Civil Rights Act of 1964. Both opinions embodied rigorous standards of test validation and seemed to signal the Court's intention to require increasingly careful study of the relationship between an employment test and the job for which it was used. These decisions also seemed to put a large number of issues from a relatively specialized area—psychological test theory—squarely before the courts. Employers and their psychologists were given notice that "professionally developed" tests were not enough: the courts would actually examine the manner in which these professionals developed and used their tests. Indeed, the Supreme Court appeared willing—and able—to hold the practitioners of psychological testing to the highest standards of their profession.

**C. Washington v. Davis: The Equal Protection Standard**

Trends begun in *Griggs* and *Albemarle* appeared to be dramatically reversed in *Washington v. Davis*. Here the Court was faced with a somewhat different issue, and returned a significantly different result.

---

83. *Id.* at 434.
84. AMERICAN PSYCHOLOGICAL ASSOCIATION, STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTS AND MANUALS (1966) [hereinafter cited as 1966 APA STANDARDS].
85. 422 U.S. at 435. A differential validity study compares separate results reported for different, similarly situated racial groups. The Court recognized that given the history of discrimination in upper-level positions, a minority test group for those positions would not be available. The Court emphasized, however, that no such justification could be offered concerning lower-level jobs. *Id.*
86. *Id.* at 436.
Explicitly abandoning the "more rigorous" statutory standard of Title VII that involved a "more probing judicial review," the Court altered both the range and domain of inquiry for employment testing cases brought on equal protection grounds.

The case involved a general test of verbal ability that was used to screen applicants to the District of Columbia police department. Two black applicants who had failed the test, and who were denied employment with the department as a result, brought suit. They challenged the department's use of "Test 21" on constitutional grounds as a violation of the Fifth Amendment, and on separate statutory grounds under both § 1981 and the District of Columbia Code. No Title VII claims were asserted.

At trial, plaintiffs introduced evidence that between 1968 and 1971 black applicants in the District of Columbia had failed Test 21 at a rate more than four times the rate for white applicants. The district court judge accepted the showing of disparate impact but, applying a Title VII standard, found that the test was "job-related." He ruled that the defendants had met their burden of demonstrating job-relatedness by showing that the test was "directly related to the requirements of the police recruit training program." The Court of Appeals, however, was not similarly persuaded and it reversed, stating: "We think this

88. Id. at 247.
89. Id. at 232.
90. Id. at 232-33.
91. The due process clause of the fifth amendment has been held to prohibit the United States from invidious discrimination between individuals or groups. Bolling v. Sharpe, 347 U.S. 497 (1954).
92. 42 U.S.C. 1981 (1976) provides:
All persons within the jurisdiction of the United States shall have the same right in every State and Territory to make and enforce contracts, to sue, be parties, give evidence, and to the full and equal benefit of all laws and proceedings for the security of persons and property as is enjoyed by white citizens, and shall be subject to like punishment, pains, penalties, taxes, licenses, and exactions of every kind, and to no other.
93. D.C. CODE 1-320 (1973) provides:
In any program of recruitment or hiring of individuals to fill positions in the government of the District of Columbia, no officer or employee of the government of the District of Columbia shall exclude or give preference to the residents of the District of Columbia or any State of the United States on the basis of residence, religion, race, color, or national origin.
Quoted in 426 U.S. at 233 n.2.
95. Id. at 233. "Test 21 was a written examination 'designed to test verbal ability, vocabulary, reading, and comprehension.'" Id. at 235 (quoting the district court opinion, 348 F. Supp. 15, 16 (D.D.C. 1972)). The judge ruled that the test did not appear to be culturally slanted in favor of whites and that the demonstration of a rational relation to job qualifications was a sufficient answer to allegations of discrimination based on disproportionate effect and lack of job performance validation. 426 U.S. at 235-36.
evidence tends to prove nothing more than that a written aptitude test will accurately predict performance on a second round of written examinations, and nothing to counter this hypothesis has been presented to us." Both courts had employed the framework of Title VII in reaching opposite results.

Justice White's majority opinion began by drawing a distinction that had not been made by the two lower courts in the case, counsel for either party, and at least nine other courts that had considered similar cases. The employment testing standards under the Constitution were different from and more lenient than the ones applied under Title VII. The clearest and most important difference pertained to the nature of the prima facie case required under each, and here the Court was emphatic: "We have never held that the constitutional standard for adjudicating claims of invidious racial discrimination is identical to the standards applicable under Title VII, and we decline to do so today." Noting that disproportionate impact, while not irrelevant, "is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution," the Court required that a racially discriminatory "purpose" be shown before "a law or other official act" would be held constitutional. In words that seemed to ignore the entire historical context of testing, the social reality of widespread racial disparities in test performance, and years of test litigation under Title VII, the Court pleaded "difficulty understanding how a law establishing a racially neutral qualification for employment is nevertheless racially discriminatory and denies 'any person . . . equal protection of the laws' simply because a greater proportion of Negroes fail to qualify than members of other racial or ethnic groups."

Davis at once distinguished equal protection from Title VII causes of action in employment testing, and placed a nearly insurmountable burden upon plaintiffs who would proceed on constitutional grounds. Moreover, it provided little guidance to future plaintiffs who would assume this burden of attempting to demonstrate discriminatory purpose in testing practice. In fact, in practical terms, Davis likely foreclosed

97. 512 F.2d 956 962 (D.C. Cir. 1975).
98. See 348 F. Supp. 15 and 512 F.2d at 957 n.2.
99. See 426 U.S. at 238 n.10, regarding the confusion at oral argument about whether Title VII or constitutional claims were at issue.
100. The Court listed sixteen lower federal court cases, nine of which involved public employment, whose contrary decisions it overruled. Id. at 244-45 n.12.
101. Id. at 239.
102. Id. at 242.
103. Id. at 245.
104. In Village of Arlington Heights v. Metropolitan Hous. Dev. Corp., 429 U.S. 252 (1977), the Court confirmed the basic rule of Davis—that "[p]roof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause," id. at 265—and undertook a brief discussion of how such proof might be accomplished. Justice Powell suggested that
any successful constitutional challenges to discriminatory employment testing. In spite of the remedies available under Title VII, this foreclosure is not without consequence. For plaintiffs who seek more than two years back pay, request punitive or compensatory damages, desire a jury trial, or wish to avoid the elaborate procedural requirements of Title VII, the constitutional standard is critically important. It is similarly important for plaintiffs not covered under Title VII. But the real impact of Davis may lie elsewhere in the opinion, in a cryptic discussion of job relatedness that has uncertain implications for Title VII itself.

D. Davis and Title VII: An Uncertain Connection

If the Court was clear and emphatic in distinguishing the nature of prima facie testing cases under Title VII and the Constitution, it was anything but that with respect to how the two differed, if at all, on standards of demonstrating job relatedness. Confusion stems from the majority's assertion that the plaintiffs in Davis were not entitled to relief on constitutional or statutory grounds. Since the plaintiffs had evidence might be found in the historical background of the decision or the sequence of events leading up to it, as well as in departures from "normal" procedures or substantive grounds for decision-making, and in the legislative or administrative history of the decision. Id. at 267-68. Few if any such "evidentiary sources" are present in the typical test discrimination cases. But see infra note 123. 105. Title VII limits relief to no more than two years back pay [42 U.S.C. § 2000e-5(g) (1976)], does not provide for punitive or compensatory damages [see Bernstein, Damages for Federal Employment Discrimination: Section 1981 and Qualified Executive Immunity, 85 YALE L.J. 518, 519-23 (1976)] or a jury trial, and imposes an elaborate set of procedural requirements that must be followed before bringing suit, 42 U.S.C. § 2000e-5 (1976).

106. Title VII defines an "employer" as a person "engaged in an industry affecting commerce who has fifteen or more employees." 42 U.S.C. § 2000e(b) (1976). Thus, employees of small businesses are not covered by the statute.

107. The statutory discussion seemed gratuitous, as Justice Brennan observed in dissent. 426 U.S. at 257. The majority opinion had conceded that defendants' motion for summary judgment, on which they had prevailed in district court, presented "[n]o issue under any statute or regulation," id. at 234, but felt compelled to discuss statutory issues nonetheless.

One statutory issue it did not address, however, was whether the Davis discriminatory purpose requirement was intended to apply to actions under section 1981. This issue continued to be actively debated. See Manishin, Section 1981: Discriminatory Purpose or Discriminatory Impact?, 80 COLUM. L. REV. 137 (1980):

[T]he vast majority of courts—typically upon summary consideration—have held that Washington v. Davis mandates use of a discriminatory purpose standard under the statute. On the infrequent occasions where courts have attempted to analyze the problems in a comprehensive manner they have arrived at opposite and irreconcilable conclusions. Id. at 138-39 (footnotes omitted). For the confusion that attends this matter, see Williams v. DeKalb County, 577 F.2d 248 (5th Cir. 1978), where the Fifth Circuit remanded a case to the district court with instructions to follow the teachings of Albemarle and Griggs in deciding educational testing issues. One judge concurred specially, urging that the case be decided on remand by the standards set forth in Davis, even though the initial trial court decision had been issued prior to Davis. Id. at 257 (Clark, J., specially concurring). Just two and one-half months later, the same three-member panel modified its opinion to conform to the concurrence and ruled that Davis was the controlling precedent, since a claim under section 1981 is to be equated with a four-
made the necessary statutory showing of disparate impact, the opinion
turned to a discussion of whether the defendants had rebutted this
showing by demonstrating job relatedness. The Court proceeded to
endorse a standard far more lenient than anything suggested by Griggs
or Albemarle: a “positive relationship between the test and training-
course performance was sufficient to validate the [test] wholly aside
from its possible relationship to actual performance as a police of-
ficer.” Insisting that this standard was not precluded by either of the
two earlier cases, the Court suggested that the new interpretation
“seems to us the much more sensible construction of the job-related-
ness requirement.” Through indirection and ambiguity, the Court
seemed to foreshadow an important change in Title VII employment
test standards.

Test advocates heralded the decision as indicative of a new atti-
dute on the part of some members of the Court, one far more lenient
towards the vagaries of employment testing than earlier. One com-
mentator suggested that “[a]n excessively heavy burden upon the test
user to prove job-relatedness . . . has been lightened substantially.”
Others saw it as the beginning of a “new trend in court decisions” that
would give employers far more latitude to deviate from EEOC
guidelines than had been applied under Title VII. Supporters of the
rigorous EEOC standards, however, lamented the decision for precisely
the same reason. One speculated that Davis may signal “the end of the
job-relatedness requirement as a strong and meaningful bar to employ-
teenth amendment, not a Title VII, claim. 582 F.2d 2, 2-3 (5th Cir. 1978). Another example of
this confusion can be found in the Ninth Circuit. Compare Davis v. County of Los Angeles, 566
F.2d 1334, 1340 (9th Cir. 1977), vacated as moot, 440 U.S. 625 (1979) (reading Washington v. Davis
to be consistent with prior case law dealing with proof of discrimination under section 1981 or
Title VII and requiring only a showing of disparate impact) with Craig v. County of Los Angeles,
626 F.2d 659, 662-63 (9th Cir. 1980), cert. denied, 450 U.S. 919 (1981) (proof of discriminatory
intent required for section 1981 and Title VII claims after Washington v. Davis). Recently, how-
ever, the situation seems to have been clarified. In General Bldg. Contractors Ass’n, Inc. v. Penn-
sylvania, 102 S. Ct. 3141 (1982), the Supreme Court held that section 1981 “can be violated only
by purposeful discrimination.” Id. at 3150.

108. 426 U.S. at 250.
109. Id
110. Id at 251.
111. For example, the Court noted that the defendants had never disputed “that under the
statutes and regulations governing their conduct standards similar to those obtaining under Title
VII had to be satisfied.” Id. at 249 (footnote omitted). The majority opinion did not dispute it
either, leading to the obvious inference that the Court was applying a standard “similar” to Title
VII in endorsing the defendants’ demonstration of job-relatedness.
112. E.g., Lerner, Washington v. Davis: Quantity, Quality, and Equality in Employment Testing,
Public With the Problems, 10 URB. LAW. 1 (1978).
113. Holtzman, Validity and Legality, in EDUCATIONAL TESTING SERVICE, EDUCATIONAL
114. Portwood & Schmidt, Beyond Griggs v. Duke Power Company: Title VII After Wash-
ment discrimination," and another suggested that "the Court in Washington v. Davis dropped any pretense of strictness with respect to job-relatedness and simultaneously abandoned its posture of deference to the EEOC. . . ."

Despite these dire predictions, the lower federal courts have not used Davis to eviscerate the rigorous standards of the earlier Title VII cases. Some district courts have applied the Davis standards to Title VII cases, both in requiring a showing of discriminatory purpose and in applying its lenient standard of job-relatedness. Respective circuit courts, however, have been uniformly unwilling to follow this lead.

The Ninth Circuit, for example, twice rejected the applicability of Davis’ job-relatedness standard to Title VII, and on one of those occasions engaged in a scrupulous review of employment test methodology fully in the spirit of both Griggs and Albemarle. In the course of that review, the court pointed to the perils of Davis’ substitution of training-course for job-relatedness: "If employers were permitted to validate selection devices without reference to job performance, then non-job-related selection devices could always be validated through the simple expedient of employing them at both the pre-training and training stage."

The Supreme Court has not yet fully examined the question of job-relatedness. However, its summary consideration of National Education Association v. South Carolina suggests that the matter is far from settled. In that case, a three-judge district court panel had approved the use of National Teacher Examinations for the certification


118. Scott v. City of Anniston, 597 F.2d 897 (5th Cir. 1979), cert. denied 446 U.S. 917 (1980) [modifying 430 F. Supp. 508 (N.D. Ala. 1977)]; Blake v. City of Los Angeles, 595 F.2d 1367 (9th Cir. 1979), cert. denied, 446 U.S. 928 (1980) [reversing 435 F. Supp. 55 (C.D. Cal. 1977)]. See also Firefighters Inst. for Racial Equality v. City of St. Louis, 616 F.2d 350 (8th Cir. 1980), cert. denied, 452 U.S. 938 (1981). In United States v. City of Chicago, 549 F.2d 415 (7th Cir.), cert. denied, 434 U.S. 875 (1977), the Seventh Circuit did not explicitly reject Davis—noting that "the applicability of the Court’s holding to cases arising under Title VII is not clear,” id. at 431—but it employed the Griggs-Albemarle standards in upholding the district court. Id. at 427.

119. Craig v. County of Los Angeles, 626 F.2d 659, 663 (9th Cir. 1980), cert. denied, 450 U.S. 919 (1981); Blake v. City of Los Angeles, 595 F.2d at 1372-74.

120. 595 F.2d 1367.

121. Id. at 1382 n.17.

and compensation of teachers. Disparate impact was not only clearly demonstrated for the current use of the test, but also well-documented for the more than thirty years that the test had been employed by the state.\textsuperscript{123} The district court found the test job-related on the basis of a validation study whose highly questionable methodology did not even attempt to relate the test to actual job performance.\textsuperscript{124} Although the case was brought under Title VII, the three-judge panel derived its job-relatedness standard entirely from \textit{Davis}.\textsuperscript{125} The Supreme Court affirmed without comment.\textsuperscript{126}

It is not yet clear what \textit{Davis}' ultimate effect will be on Title VII cases. What is clear, however, is that the decision virtually insures continuing debate over the dimensions of a fair yet practical policy of employment testing. At the same time, there is virtually no evidence that the widespread practices that gave rise to decisions like \textit{Griggs} and \textit{Albemarle} have ended. Few, if any, of the methodological problems have been resolved, and no real innovations in traditional testing practices have been developed and implemented. Indeed, there is now more reason to question the use of these techniques than in the past.\textsuperscript{127} The following pages present a brief but critical assessment of the inadequacies of standard employment test practices. It is offered in the belief that a putatively "liberalized" attitude on the part of the Court is a step decidedly in the wrong direction. And it is intended in part as a coun-

\textsuperscript{123} South Carolina's use of the National Teacher Examinations provides a suggestive example of how "neutral" practices can produce racially discriminatory effects on a widespread basis. The state had traditionally maintained a "dual" school system that racially segregated students and teachers. Black teachers with the same responsibilities as their white counterparts were paid less. In 1940, however, the Fourth Circuit held that a similar dual pay system in Virginia was unconstitutional. Alston v. School Board, 112 F.2d 992 (4th Cir.), cert. denied, 311 U.S. 693 (1940). South Carolina's legislature then appointed a committee to examine its own dual pay system. The committee returned with a recommendation that the system be unified but that teacher salaries be pegged to scores on the National Teacher Examinations which, it noted, could be "scored objectively and impartially and their use would not be subject to the accusation that they are used for purposes of discrimination." \textit{Quoted in} 445 F. Supp. at 1102. When the first half of the teachers in the state were tested, the results showed that 90% of the white teachers would qualify for the two highest paying certificates, while 73% of the black teachers would be relegated to the two lowest paying categories. \textit{Id} at 1102. The tests were formally adopted. Apparently, the state could not bring itself to pay black and white teachers equally, perhaps in part because it could not conceive of the black teachers as being as competent as the whites. The tests solved both problems.

\textsuperscript{124} \textit{See infra} notes 216-220 and accompanying text.

\textsuperscript{125} 445 F. Supp. at 1113. The court provided this explanation:

\textit{The Supreme Court made clear once again in} \textit{Washington v. Davis} that a content validity study that satisfies professional standards also satisfies Title VII. . . . \textit{[T]he decision to validate against the academic training program rather than job performance is specifically endorsed in principle in} \textit{Davis} . . . .

\textit{Id.} (citing \textit{Davis}, 426 U.S. at 247 n.13).

\textsuperscript{126} 434 U.S. 1026 (1978).

\textsuperscript{127} \textit{See infra} Part IV.
terbalance to the relatively one-sided professional opinion marshalled in support of employment testing.

III. STATE OF THE ART: ON THE NATURE AND CONSEQUENCES OF EMPLOYMENT TESTING

Employment testing now has the appearance of a highly specialized and technical industry, increasingly dominated by elaborate statistical formulae, sophisticated selection models, and complicated research designs. One federal judge, for example, has complained that "Title VII cases are contests between college professor statisticians who revel in discoursing about advanced statistical theory."128 Like all technologies, however, employment testing cannot transcend the limits of its basic assumptions and its essential methods. In this section, I will examine some of the assumptions and methods that are too often ignored in test industry attempts to attain increasing levels of apparent sophistication. These issues remain at the center of legal debates over fair employment test policy. One test theorist has suggested that research on testing has become "forensic as well as investigative."129 Indeed, no forensically-motivated examination of employment testing should ignore these basic features of the enterprise.

The apparent precision and objectivity of employment tests may be their most seductive qualities. Each administration of these tests produces clean, precise numbers that appear to indicate something important about those who take them, and to simplify enormously the hiring and promotion decisions faced by employers. Employment tests undoubtedly simplify things. Whether they produce numbers that have any real meaning, however, is an entirely open empirical question—an empirical question that is too infrequently answered in a meaningful way. This question should become more focused when considered in the following context.

A. Disparate Impact: The Social Reality of Testing

The first and most important point to be made about these tests is that they do adversely affect minorities. Numerous testing cases brought under Title VII attest to this.130 One consistent finding that

130. The cases ordinarily employ a "four-fifths" rule of thumb in determining adverse impact. That is, a testing procedure is not ordinarily regarded as producing disparate impact unless the minority pass rate is less than four-fifths the rate for the highest group. See 43 Fed. Reg. 38,297-98 (1978). See also supra note 76.
emerges from the psychological literature on testing is that minorities tend to do poorly on such standardized measures, at least as compared to their white counterparts.\footnote{131} The staunchest advocates of the tests concede this. For example, one of them writes of the plaintiff's Title VII burden to demonstrate disparate impact that "since tests in general have this effect, producing such evidence is no more difficult than picking up stones from a gravel road."\footnote{132} Despite the ubiquity of such


When standard aptitude tests are given to low income Negro children, or disadvantaged children . . . the tests are less precise and less accurate—so much so that test scores become practically meaningless. Because of the impoverished circumstances that characterize the disadvantaged child, it is virtually impossible to tell whether the test score reflects lack of ability—or simply lack of opportunity. Moreover, the probability that test scores of the Negro child or the disadvantaged child will be depressed because of somewhat unique psychological influences further compounds the risk of inaccuracy.

269 F. Supp. at 485. Judge Wright cited, inter alia, the Lorton Study in which the results of nonverbal I.Q. tests given "educationally deprived" children were an average of 29 points higher than those from standard verbal I.Q. tests. See N. Burke & A. Simons, A Measure of the Educational Achievement of a Group of Incarcerated Culturally Disadvantaged and Educationally Deprived Dropouts (1965).

132. Seelman, supra note 59, at 55.

A distinction is often made between differential validity [a test may correlate with performance measures less well for some groups than others] and test bias [a test may systematically over or under predict the performance of certain groups compared to others]. "Fairness" is what is ordinarily said to be at issue in the legal examination of these test characteristics. It would be unfair, for example, to exclude some people from employment on the basis of scores on a test that was valid only for another group of people. A test might be "valid" but also biased. That is, there might be an equally strong relationship between test scores and performance measures, but because of the test bias, cutoff scores for certain groups would be set at levels systematically too high or too low.

evidence, members of the employment test industry were strangely silent on the issue until legal regulation attracted their attention to it. *Annual Review of Psychology* chapters on "industrial psychology," "personnel selection," and "personnel management," commencing in 1950, did not even mention disparate impact or racial differences in employment test performance until 1965, and then only in the context of the cross-cultural use of the tests. Not until two years later was the possibility that employment tests might be less fair for minority than for white job applicants given serious consideration.

Thus, although employment testing is nearly a century old, serious industry consideration of disparate impact is of relatively recent origin. This fact may help to explain why the employment test specialists have successfully resisted the clear implication of disparate impact statistics: the more such tests are used, the more importance is attached to them in employment decisions, and the more minorities will be disadvantaged in the workplace.

Locating the ultimate source of this disparity in test performance, however, is highly controversial. Some of the deficit can be explained by the nature of the tests themselves and the conditions under which they are administered. For example, since most test administrators are white, they may be unable to develop adequate rapport with minority applicants. The lack of rapport may adversely affect test performance. Most tests are administered in academic settings—timed exams, paper and pencil tests, administration by white authority figures, and so on. Minorities may be alienated or intimidated by the setting and do poorly on the test as a result. Other problems relate less to the test situation but more to what is being measured. Many tests incorporate the common knowledge and assumptions of the dominant culture, and implicitly exclude minority perspectives. Moreover, most employment tests are administered verbally, either orally or in writing. Minorities traditionally perform poorly on standardized measures of verbal skills. The cases speak as clearly to this point as the psychological

---

137. See supra note 131.
research does. In *Scott v. City of Anniston*,\(^{138}\) for example, only four of seventeen blacks who were given a written employment test passed, as opposed to all twenty-two whites who took it. Some of the blacks who failed had been successfully performing the job in question and were demoted as a result of their failure, while other blacks who failed the written test passed a performance test given at a later time. Put simply, to the extent that performance measures of non-verbal competencies depend in part on verbal ability, the test is measuring the wrong thing, and it does so to the detriment of the minorities who take it.

Some performance disparities are undoubtedly the consequence of the conditions of segregation, discrimination, and deprivation to which minorities have been subjected.\(^{139}\) That is, they may be "real." Inferior schools, poor educational opportunity, and discrimination in the workplace may have taken their toll among minorities subjected to them.\(^{140}\) We have no way of knowing, however, how much of the disparity in test performance is due to "real" differences in skills. Nor do we know how much of this difference is remediable through short-term training programs. Despite their shortcomings, employment tests produce seemingly scientific results that make all differences appear real. By cloaking imperfect measurements beneath the mantle of "science," the numbers are given an air of authority, legitimacy, and "truth" that they would not otherwise have. Techniques that are imprecise and error-ridden produce numbers that appear "objective" and true. Thus, the same author who compared locating evidence of disparate impact to "picking up stones from a gravel road," can state, just two pages later:

Tests are, however, inherently neutral selection procedures. Moreover, they can discriminate on the basis of merit and other non-prohibited grounds. Consequently, one cannot assume that tests are inherently destructive to any group.\(^{141}\)

Whether disparities in test results are actual or artifactual is in some ways less important than the meaning that is attached to them. If employment test scores are regarded as nothing more than estimates of test-taking ability, then the ultimate "reality" of performance differences is of far less consequence than if scores are taken to be important predictors of future job performance. But test numbers not only appear objective and neutral, they are also rendered tangible and personal in interpretation. "Test constructors and users as we have known them

---

138. 597 F.2d 897 (5th Cir. 1979), *cert. denied*, 446 U.S. 917 (1980).
140. But there is no convincing evidence that racial disparities have their origins in anything but disparities in living conditions and treatment. See L. Kamin, *supra* note 5; Block & Dworkin, *I.Q., Heritability, and Inequality* (pt. 1), 3 *Phil. & Pub. Aff.* 331 (1974).
have generally been prone to reifying and hypostatizing, prone to assume that tests were tapping dispositional syndromes with other symptoms than those utilized in the test."142 Thus, disparities in test performance are translated into disparities in personal characteristics like ability, intelligence, and merit with implications far beyond the test itself. The tendency is not confined to test developers and users. In National Education Association,143 for example, the district court was presented with National Teacher Examinations that were designed to measure nothing more than retention of college course material in various areas of academic specialization. No attempt whatsoever had been made in these tests to assess teaching performance, or to relate the test scores in any way to teaching skill. Nonetheless, the court insisted on referring to persons who scored in the lower (but still acceptable) range on the tests as “without adequate knowledge to teach effectively,” “minimally competent,” and finally, “unqualified teachers.”144

The characteristics supposedly measured by tests are often seen as “relatively permanent, underlying, causal entities.”145 Test developers often disclaim this causal implication. One writes that “[n]o test can eliminate causality. Nor can a test score, however derived, reveal the origin of the behavior it reflects.”146 But the disclaimer is misleading, since the test does purport to measure one, and only one, causal entity—the individual. In this sense, tests encourage and promote what social psychologists call “fundamental attribution error”—accounting for behavior in terms of the traits or dispositions of the people who perform it rather than the characteristics of the situation in which it occurs.147 Test factors that compromise the performance of minority applicants are discounted in favor of explanations that locate poor performance in poor ability and the absence of talent. The ultimate origins of disparate impact are still attributed by some test advocates primarily to heredity and genetic deficiency.148 But most enlightened test developers and users concede the environmental basis for performance differences, which they regard as real differences in ability none-

142. Campbell, Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity, 15 AM. PSYCHOLOGIST 546 (1960).
144. Id. at 1116.
148. One of the best known contemporary proponents of doctrines of racial genetic inferiority has recently concerned himself, at length, with test methodology. See A. JENSEN, BIAS IN MENTAL TESTING (1980). In the course of concluding that mental tests are relatively without fault, he includes this ironic, acausal caveat: “[T]he constructors, publishers, and users of tests are under no obligation to explain the causes of the statistical differences in test scores between various subpopulations. They can remain agnostic on that issue.” Id. at 740 (emphasis in original).
theless. Here too, though, there is often a permanency ascribed to traits and dispositions supposedly measured by the tests. "Environmentally caused differences can be as pervasive, enduring, and basic in the life of the individual as those caused by heredity."149 Whether the product of genetic deficiency or environmental deprivation, low test scores are easily taken to represent relatively stable properties of the person "measured."

Disparate impact thus represents the apparent objectification of racial differences in ability and skill. Not surprisingly, when test scores are treated as indices of merit and worth by test administrators and employers, the results are often internalized by test-takers as well. Indeed, the tests encourage this form of internalization, appearing to represent unbiased and sophisticated documentation of inadequacy and personal shortcoming for which one has only oneself to blame.150 Despite the disclaimers about "causal inference," the social reality of disparate impact is that minorities are taught, in disproportionate numbers, that they do not "measure up."

One outspoken advocate defends tests against the claim of racism by observing that the "overprivileged white opponents of testing outnumber underprivileged black ones and always have, not only in absolute numbers but in relative, proportional terms as well."151 If the statement is correct, it may underscore testing's most pernicious aspect—the ease with which tests convince those who fail them that it is not the tests, but themselves, who are in need of fixing. Courts have just begun to recognize the psychological as well as economic components to disparate impact, components that make test discrimination in some ways more cruel than the overt racial prejudices of an earlier time.152

149. Anastasi, supra note 146, at 392.
To make the point even more vividly, suppose you are a ghetto resident in the Roxbury section of Boston. To qualify for being a policeman you have to take a three-hour-long general intelligence test in which you must know the meaning of words like "quell," "pyromaniac," and "lexicon." If you do not know enough of those words or cannot play analogy games with them, you do not qualify and must be satisfied with some such job as being a janitor for which an "intelligence" test is not required yet by the Massachusetts Civil Service Commission. You, not unreasonably, feel angry, upset, and unsuccessful. Because you do not know those words, you are considered to have low intelligence, and since you consequently have to take a low-status job and are unhappy, you contribute to the celebrated correlations of low intelligence with low occupational status and poor adjustment. Psychologists should be ashamed of themselves for promoting a view of general intelligence that has encouraged such a testing program, particularly when there is no solid evidence that significantly relates performance on this type of intelligence tests with performance as a policeman.

B. Job Relatedness and Job Analysis: What Do Employment Tests Measure?

Test advocates often suggest that public criticism and the legal regulation of testing is motivated by a desire to reject the “message” of testing, including the message of racial differences in performance. “In the employment realm,” one of them writes, “the problem with school-like paper and pencil tests is not that they do not work; the problem is that too many of them work too well” by telling us “unpleasant and unflattering truths” about ourselves and our society. Legal scrutiny of tests that register disparate impact is thus likened to “a modern version of killing the messenger who brings bad news.” When viewed in the context of the legal standards that have actually been applied to testing, however, this analogy is seriously misconceived. At their most rigorous, all that Title VII cases have ever required was that when, in fact, the news was “bad,” the messenger demonstrate the test’s accuracy and relevance.

For the bad news of disparate impact, the courts have required that testers demonstrate the job relatedness of test scores. Despite the protests of test advocates, demands for such evidence do not pose a serious threat to the meritocratic values that tests supposedly represent. Indeed, in one sense, such demands are merely an extension of these values. As the Chairman of the Civil Service Commission once observed, the “necessity to measure characteristics of people that are related to job performance is at the very root of the merit system.” Thus, the real question becomes whether differences in test performance tell us anything important about job performance. Much evidence suggests that they do not. Thorndike and Hagen, for example, examined over 12,000 correlations between test scores and subsequent

The stigma resulting from failure on an invalid examination is both an individual emotional trauma and an apparently objective badge of inferiority inseparably linked to the failing examinee’s race or national origin, thus going to the very heart of Title VII, 42 U.S.C. § 1981, and the fourteenth amendment. The fact that the examination may later be declared invalid and the injured persons accorded an opportunity to take a valid test cannot erase the stigma: the fact will remain that a disproportionate number of minority candidates failed an examination that a disproportionate number of white males passed . . . . [T]he social and emotional harm flowing from an invalid promotional examination will remain beyond the time and opportunity to take a valid examination is afforded.

153. Lerner, supra note 151, at 11-12 (emphasis added).
To conceal the effects of cultural disadvantages by rejecting tests or by trying to devise tests that are insensitive to such effects can only retard progress toward a genuine solution of social problems. Such reactions toward tests are equivalent to breaking a thermometer because it registers a body temperature of 101°.
155. E.g., Lerner, supra note 112; Seelman, supra note 59. See also supra note 71 and references cited therein.
job performance from data obtained from over 10,000 respondents. The number of significant correlations did not exceed the number expected as a result of chance. Attempts at test validation have been most impressive when tests were used to predict subsequent test-taking behavior rather than actual job performance. As one moves farther away from test-like situations, the statistics become increasingly less impressive. The point is not that test scores are never related to job behavior, but that it would be very foolish to assume that they are. Thus, when courts have required a showing of "job relatedness," they have been asking for nothing more than a much needed demonstration that the particular test measures something important about the applicant's capacity to do the job in question.

One commentator writes that:

[j]ests provide three distinct advantages over less disciplined alternatives: they are standardized, providing every examinee with the same instructions, the same tasks to perform, and the same time limits; they are efficient, enabling the collection of a considerable amount of information about a person in a short amount of time; and they are amenable to objective analysis, furnishing test results which may be quantified and subjected to statistical evaluation.

Indeed, these are the reasons fair minded employers use tests and judges approve them in spite of their disparate impact. But these "advantages" are meaningless in the absence of job relatedness. Moreover, such advantages cannot be asserted, but must be demonstrated instead. "Standardization," for example, is a questionable assumption when one realizes the different perspectives that potential employees bring to the tests. Identical instructions do not mean that all people understand them in the same way. Similarly, employment tests are "efficient" inasmuch as they enable rapid collection of data. Whether those data can be regarded as information (i.e., are informative), however, is another matter, one that depends entirely on the demonstration of job relatedness. Hence, "efficiency" cannot justify the use of a test in the absence of such a demonstration. Finally, the quantified, "objective analysis" of these data is valuable only if the data are meaningful. The objective analysis of biased or invalid numbers is, at best, useless.

It is also true that a demonstration of job relatedness "merely ensures that the selection device chooses a successful employee with significantly greater frequency than could be expected from pure

157. R. THORNDIKE & E. HAGEN, 10,000 CAREERS (1959). See also Lent, Aurbach, & Levin, Predictors, Criteria, and Significant Results, 24 Personnel Psychology 519 (1971), who found that less than half of the 1506 reported test-job performance relationships they reviewed were statistically significant, and E. GHISELLI, THE VALIDITY OF OCCUPATIONAL APTITUDE TESTS (1966), who reported statistical relationships in which test performance typically accounted for less than ten percent of the variance in job performance.

158. Johnson, supra note 72, at 1257.
chance." But "chance" is not necessarily the most relevant comparison group here. It is likely that a variety of selection procedures—including those that do not use any tests at all—are successful beyond chance in selecting successful employees. Particularly when the test results in differential minority hiring and promotion, the more important issue becomes whether the test can select successfully with significantly greater frequency than available alternatives. If it cannot, and those alternatives are neither prohibitively expensive nor themselves discriminatory, there is no apparent reason why the original test should continue to be used. Arguments to the effect that requiring demonstrations of job relatedness will lead to the use of more subjective and discriminatory selection procedures fail to acknowledge the applicability of disparate impact analysis to all forms of employment selection. Subjective procedures that adversely impact minorities and are not job related are, of course, equally impermissible.

Moreover, job relatedness does not mean training program relatedness. In Washington v. Davis, the Supreme Court based its finding of job relatedness on a validation study that used only training school performance as its criterion. Because training school and job performance may or may not be related, the Court erred if it intended to equate them. Indeed, in Davis, training school performance apparently predicted job performance only for white police officers, not for blacks. Thus, training school performance may tell employers absolutely nothing about job performance, and it is impossible to know until the appropriate empirical study is done. It is difficult to justify denying employment to people on the basis of test scores, the only significance of which is their ability to predict performance in what may be an inadequate or largely irrelevant training program.

Training program performance may not only be a poor predictor of job performance, it may be systematically biased as well. Training programs tend to resemble academic settings in style and content, and employ methods of assessment that are written and standardized. There may be a large overlap in the measurement error in both settings—common but irrelevant factors which affect performance—that accounts for any correlation between them. Since some minorities tend to perform poorly in formal academic settings, test scores may "validly" predict poor training program performance for minorities but at the same time seriously underpredict their performance in the less aca-

159. Id. at 1259.
160. E.g., Lerner, supra note 112, at 306.
162. 426 U.S. at 251 n.17. See Justice Brennan’s discussion of this issue beginning at 259.
163. Id. at 267 n.10. But see 512 F.2d at 963 n.44.
EMPLOYMENT TESTS AND DISCRIMINATION

A testing policy like the one approved in Davis, that permits validation on the basis of training performance criteria, incorporates this bias and may significantly disadvantage minorities as a result.

Meaningful studies of the relationship of test scores to job performance require a careful analysis of the nature of the job itself. Thus, determinations of job relatedness depend heavily upon job analysis. Yet, at this crucial initial stage, the technology of testing stands relatively undeveloped. One specialist in the area concludes that in spite of considerable research on the topic, "the degree of clarity and the extent of our knowledge about what we are doing and what we have accomplished leaves considerable [sic] to be desired." Characterizing earlier research as "little more than a pilot effort, particularly with reference to personnel selection," he suggests that "[d]uring the past five years, additional research results have accumulated, but only a few studies appear to advance the state-of-the-art.

Concern over job analysis predates litigation under Title VII. Not until recently, however, have courts and psychologists recognized the logical centrality of job analysis to legally acceptable demonstrations of job relatedness. Indeed, one review of thirty-one federal testing cases found that job analysis had been conducted by the defendants for purposes of test development in only eleven of them. Moreover, failure to conduct the analysis appeared to have little consequence for the outcome of the cases—in only three cases was the court's ruling based on the absence of job analysis.

Job analysis is not always done, and when it is, it is not always done well. The main problem is that job analysis depends entirely upon opinion. Someone—a psychologist, a supervisor, employees, experts—must offer an opinion about what is important or essential in effectively performing the job in question. These opinions can be more or less systematically evaluated and combined, but the job analysis cannot transcend the perspectives of those who participate in it. If the opinions are ill-considered, incomplete, or incorrect, or if they incorpo-
rate outside biases and stereotypes, the job analysis will bear these marks. No amount of psychometric sophistication or elegant statistical analysis added after the fact in validation studies can correct for job analysis that is flawed or imprecise.

Careful job analysis requires the use of a "theory" about the job, rather than simply an "objective" assessment of current job practices. The more complex or ill-defined the job, the more theoretical the job analysis becomes. The particular aspects of the job that come to be regarded as "important" depend in large part on whose theory is brought to bear in the job analysis. Which theory is chosen in this analysis thus has important implications for both job relatedness and disparate impact. "Test 21," at issue in Washington v. Davis, was found by the district court judge to relate to important aspects of police job performance. In his "job analysis," he observed:

Study of the syllabus of the training course readily demonstrates the intricacy of police procedures, the emphasis on report writing, the need to differentiate elements of numerous offenses and legal rulings, and the subtleties of training required in behavioral sciences and related disciplines. Daily the significance of these skills demanding reasoning and verbal and literacy skills is borne out of the crucible of the criminal trial court.170

The judge's special concern with this aspect of a police officer's literacy skills is not surprising. On the other hand, no measure of a recruit's potential to communicate in the "crucible" of the street corner was included; consequently, none could figure in the selection process.

The relatively undeveloped nature of job analysis171 can be attributed in part to the fact that, in the past, employment testers assumed rather than investigated the job relatedness of their assessment devices. Vague and general talents and abilities, which the tests purport to measure were presumably related to the job. One salutary consequence of testing litigation under Title VII is that employment test psychologists have been forced to examine these presumptions. In the absence of such an empirical examination, job relatedness remains a crude and speculative hypothesis.

C. Predictive Validity: Checking the Employment Test Hypothesis

In employment discrimination cases, tests are ordinarily evaluated in terms of their "validity." The concept of validity represents an estimate of the degree to which tests measure what they are designed and used to measure. Validity varies across tests, of course, but also, for the same test, it varies across populations, conditions of administration,

170. 348 F. Supp. at 17. For a discussion of the same passage that provides a significantly different interpretation, see Lerner, supra note 112, at 284.
171. See Dunnette & Borman, supra note 16, at 486, for a list of "thorny methodological issues" that continue to plague job analysis techniques.
and so on. Moreover, it will also vary as a function of the particular use to which the test is being put and the kind of validity that is being measured.

The discussions of validity that have dominated court opinions and scholarly articles sometimes convey the misleading impression that "validity" is a global, tangible quality, one that some tests have and others do not. The term is employed almost in the fashion of a trait label, as though validity were some enduring property like "baldness" that attached more or less permanently to the test. Moreover, disputes about types of validity—ubiquitous in case law involving employment tests—are often portrayed as exercises in academic esoterica. Once a test has been demonstrated "valid," some advocates have argued, that should be enough. But the different kinds of validity are not just different levels in a hierarchy of increasing rigor. Instead, they tell test developers very different things about the usefulness of the test in question. One is not justified in assuming that, because a test is content or construct valid, it must be "somewhat" or tolerably criterion valid.\footnote{172} The kind of validity coefficient that is appropriate or useful to calculate depends upon the use to which the test—and coefficient—is being put.

Of course, the ultimate use to which an employment test will be put is prediction: employers wish to predict who will, and who will not, perform successfully in a particular job. In albeit simplified form, this is what can be called the "employment test hypothesis"—that test scores predict job performance. It is a predictive hypothesis that can only be evaluated by a determination of predictive validity.\footnote{173} Thus, it

\footnote{172} These are the three categories of validity recognized in the 1966 APA STANDARDS, supra note 84. Content validity "is demonstrated by showing how well the content of the test samples the class (of) situations or subject matter about which conclusions are to be drawn." \textit{Id.} at 12. Construct validity estimates "the degree to which certain explanatory concepts or constructs account for performance on the test." \textit{Id.} at 13. And criterion validity is shown by "comparing the test scores with one or more external variables considered to provide a direct measure of the characteristic or behavior in question." \textit{Id.} In 1974, a much expanded revision of the Standards appeared, but these three basic categories were carried forward essentially unchanged.

\footnote{173} To emphasize the extent to which prediction is at issue, I will use this term to refer to most of what is described as criterion validity under the APA rubric. Predictive validation is empirical in nature. In studies of predictive validity the test scores of job applicants are checked statistically against actual on-the-job performance. The degree to which test scores are related to job performance scores—typically in the form of statistical correlation—provides the measure of the test's predictive validity. Selection of employees is problematic at this stage. Ideally, selection should be done on a random basis since other methods distort the distribution of test scores upon which the correlation is based. Since this is rarely deemed feasible by employers, random assignment is often dispensed with. A special distinction is often made for concurrent validation studies that compute the relationship between test scores and the performance of persons already on the job (i.e., the current work force). This fact limits the predictive utility of such studies because (1) current employees are presumably not representative of the total applicant pool, and (2) job performance and test performance correlations may be different for employees who are now more experienced than when they applied for the job in question. Despite these limitations, I will use
is measures of predictive validity that are relevant in employment testing cases, not their more convenient substitutes. All other forms of validation involve estimates of how good the test looks or seems from different perspectives. None of them, however, can evaluate the employment test hypothesis because they fail to directly incorporate empirical measures of job performance.

Yet some psychologists argue that measures of predictive validity impose too stringent a standard. At the time of Griggs, for example, one psychologist complained that "[b]oth the Supreme Court and federal agencies have seized on the logical simplicities of predictive validity without full recognition that they are imposing a standard of excellence that does not exist in other areas of decision making." Perhaps. But fairness, not logical simplicity, is at issue here. If an employer wishes to use a test that disproportionately excludes minorities, and he justifies the use of this test on the basis of his need to select employees who perform well on the job, shouldn't he be required to show—indeed, shouldn't he want to know—that the test, in fact, predicts job performance? If this is a "standard of excellence" that test developers cannot meet, then we should lament the state of the art, rather than condone the use of insufficiently job-related screening devices.

Indeed, some important limitations to the measurement of even predictive validity should be noted. "Statistical significance," the concept used to evaluate validity coefficients, is merely a convention, a matter of professional consensus. Particularly when obtained with large subject samples, statistically significant predictive validity coefficients still may not tell employers much about employees' job behavior. A predictive validity coefficient of .30, reasonably high in this context, generally means that only about ten percent of job performance can be accounted for in terms of test scores. Wider recognition that many

the term "predictive validity" to include concurrent validation since it, too, directly attempts to empirically determine the extent to which test scores can predict job performance.

Both content and construct validation, on the other hand, are non-empirical techniques, sometimes termed "rational." In both cases the test developer makes educated guesses about the extent to which the test samples the content of the job, or the underlying construct that he believes "causes" successful job performance. Unless some attempt is made to actually correlate test scores with job performance through either predictive or concurrent validation, however, there is no way to adequately evaluate the quality of those guesses, or determine how "rational" this particular approach is. See Holtzman, supra note 113, for another discussion of these concepts.

174. Fincher, supra note 73, at 494.

175. The square of the validity coefficient, $r^2_{ye}$, is termed the "coefficient of determination." First introduced in C. Hull, APTITUDE TESTING (1928), it estimates the percent of variance in actual (job) performance that can be accounted for by the test scores. For examples of how a related estimate of test source validity might vary under hypothetical conditions, see Taylor & Russell, The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables, 23 J. APPLIED PSYCHOLOGY 565 (1939). But see Brogden, On the Interpretation of the Correlation Coefficient as a Measure of Predictive Efficiency, 37 J. EDUC. PSYCHOLOGY
employment tests provide such limited increments in predictive knowledge may lead to a fuller appreciation of the mandate for their stringent legal regulation. The realities of disparate impact testing mean that these marginal increments in predictive knowledge are purchased far more at the expense of minority applicants than others. In a typical program of testing in which low validities combine with disparate impact:

there would be significant numbers of both minorities and whites for whom the predictions [would be] substantially inaccurate. For the [hired or] admitted whites, this inaccuracy would be revealed when they performed differently than the test had predicted. But for the minority applicants, most of whom would have already been rejected on the basis of their score, the opportunity to prove the test wrong—as many could statistically be expected to do, given its low validity—would never come to pass. 176

Statistically significant predictive validity coefficients also should not be confused with causal relationships. A significant coefficient does not prove that the measured ability causes the job performance; it indicates only that the two co-vary sufficiently well to use one as a rough estimate of the other. It is possible—even likely—that some third unspecified and unmeasured variable causes both. A test tends to reify the hypothesized trait or ability dimension when it may be the unmeasured moderator variable that deserves our causal attention. For example, socioeconomic status is a fairly good predictor of success on tests of verbal ability which, in turn, is a moderately good predictor of certain kinds of job success. 177 Fixation on measurement of “verbal ability” conveys the misleading impression that this trait causes success in certain occupations, rather than the equally plausible explanation that socioeconomic status (and its accoutrements) causes both.

Additional factors compromise the measurement of predictive validity. The ability of a test to predict to some job performance criterion is a function of both the test and the criterion. “A test which claims legitimacy solely on the basis of its ability to correlate with a criterion can be no better than the criterion itself which, in many instances, means no good at all.” 178 Criterion measures may be poorly defined because employers may not have considered precisely what behaviors constitute a “good job.” Of course, before any job analysis can provide

---

65 (1946). See also Dunnette & Kirchner, Validities, Vectors, and Verities, 46 J. APPLIED PSYCHOLOGY 296 (1962), for a discussion of “practical” as opposed to “statistical” significance, and Curtis, Predictive Value Compared to Predictive Validity, 26 AM. PSYCHOLOGIST 908 (1971), for a discussion of the conditions under which validity coefficients overstate predictive value.

177. See McClelland, supra note 150, especially the discussion beginning at 4.
178. Lerner, supra note 112, at 299.
the basis for a study of job relatedness, its categories must be translated into measures of performance. There is much opportunity for adequate job analysis to evolve into inadequate performance measures. "Being able to argue that the criterion is valid is as necessary as being able to show a test-criterion relationship."

Worse than being poorly defined, performance criteria may be explicitly biased. One labor scholar notes that both concurrent and predictive validation often "suffer from a fundamental defect—they are necessarily dependent on subjective criteria of supervisors and thus may be troublesome where the supervisory work force is all white." In one study of supervisory ratings, for example, researchers found that police supervisors believed black policemen were "inferior." It also showed that black patrolmen who were rated highly by their supervisors scored highest on the Defference scale of the Edwards Personal Preference Test. Stereotypic preferences play an important but largely unexamined role in predictive validation studies. Performance criteria that include supervisory approval incorporate the biases of the supervisors.

A similar concern led the Fifth Circuit in Fisher v. Proctor & Gamble Manufacturing Co. to reject a "total assessment" procedure by which employees were considered for promotion in a large manufacturing plant. The case provides a useful illustration of the subtle problems of subjectivity in employment decision-making. The company employed an elaborate process of assessment in which various kinds of information were collected about candidates for promotion into critical non-management positions. The nature of the information assembled was determined in part, by a job analysis in which the company surveyed its supervisors to determine their opinions of key skills for effective job performance. Once all of the information had been assembled, managers and supervisors in the plant rated the candidates.

179. Cronbach, supra note 129, at 48. The 1966 APA Standards also speak directly to this issue:

[T]he logic of criterion-related validity assumes that the criterion possesses validity. All too often, tests are validated against any available criterion with no corresponding investigation of the criterion itself. . . . Criterion-related validity studies based on the "criterion at hand" chosen more for availability than for a place in a carefully reasoned hypothesis, are to be deplored.


180. W. Gould, Black Workers in White Unions: Job Discrimination in the United States 156 (1977). See Lawler, The Multitrait-Multitrait Approach to Measuring Managerial Job Performance, 51 J. APPLIED PSYCHOLOGY 369, 369 (1967), to the effect that "despite impressive pleas for new approaches, the superior's evaluation is still the most frequently used measure where criteria are needed either for research purposes or for personnel decision-making purposes."


182. 613 F.2d 527 (5th Cir. 1980).

183. The information included two employment tests that the court found had not been properly validated. Id at 545.
Promotions were made from among those candidates with the highest ratings.

The court noted that no black had ever been a supervisor at the plant, and none had ever served as a rater in the assessment process itself. In this context, the “substantial subjectivity” of the assessment process was particularly worrisome, since raters were given “the opportunity to choose which events to emphasize or omit” in the ratings they assigned. The court concluded that “promotion systems utilizing subjective evaluations by all white supervisors provide a ready mechanism for discrimination.” It might well have added that studies of predictive validity whose criteria depend on such evaluations should be similarly suspect.

Despite these problems—the limited amount of added knowledge even significant coefficients provide, the possibility of inadequate or unrepresentative criterion measures, and bias in supervisory evaluations—predictive validity remains the only method of evaluating the employment test hypothesis. The possibility of flawed or inadequate validation studies provides no justification for requiring less than predictive validity as the measure of job relatedness.

D. Content Validity: When Are “Professionally Developed” Tests Good Enough?

Despite the logical centrality of predictive validity in employment testing, and the primacy afforded it by the EEOC guidelines and by the courts, many test psychologists have objected to it as unrealistic, expensive, and cumbersome. Shortly after Albemarle was decided, one lawyer told a conference of test psychologists that the opinion, and its emphasis on predictive validation, created the “real risk that aptitude testing in employment will simply fade away” because employers may decide that “the testing game is not worth the legal candle.” He advised the psychologists to “pull themselves together” and make their views felt—“[a]nd the first and foremost place to make them felt is in the EEOC Guidelines.”

In 1978, the EEOC and several other federal agencies overcame a decade of disagreement on testing policy and issued a detailed set of

184. At the time the action was filed, one of 63 management positions was held by a black, and no black had ever been a line supervisor or operations manager. Id. at 534.
185. Id. at 545 (quoting the district court).
186. Id. at 546.
187. E.g., Lerner, supra note 112; see also supra notes 71-72 and references cited therein.
189. Id. at 78.
190. Id. at 65.
191. See, e.g., Seelman, supra note 59; Note, Employment Testing and the Federal Executive
Uniform Guidelines on Employee Selection Procedures. These guidelines modified somewhat the earlier EEOC position on predictive validity; instead of expressing a blanket preference for predictive validity, a set of conditions were specified under which each of the three types of validity were deemed to be appropriate. Although a careful reading of the relevant passages does not necessarily support such an interpretation, some test psychologists have concluded that the Uniform Guidelines “see the categories as alternative, nearly independent ways to justify a selection plan.”

Coinciding with this possible shift in agency policy on the question of validation, test psychologists renewed arguments that “the very method of test construction may provide sufficient justification for test use, without further empirical investigation.” In particular, “content validation” or content sampling was—and continues to be—frequently offered as an appropriate substitute for the rigors of predictive validation. Thus, this methodology bears careful examination. One prominent test theorist who has written widely on the topic suggests that content validation does not attempt to evaluate a relationship between ability and job performance; rather it “implies at least partial identity or overlap” between the test items and the job itself. Since the test samples the job directly, there is no need to check on whether performance on both is related. He writes that eventual “performance in the domain sampled [on the job] will be reflected in performance on the sample itself [the test].”

Nothing could be more straightforward. Except that in practice several important transformations are typically effected in the job tasks as they become content samples. Thus, the degree of “overlap” becomes very much hypothesis and as much in need of empirical confirmation as the original employment test hypothesis. What makes employment tests attractive to employers is their convenience and ease
of administration. To achieve these features, however, job tasks are typically abbreviated, simplified, standardized, and in other ways transformed. These transformations may inadvertently but seriously alter the nature of the performance measure itself. The desire to reduce all tasks, however complex and non-verbal, to paper-and-pencil measurements is an expression of this concern for convenience over faithful job sampling.

In one case, for example, a fire captain's examination used multiple choice questions and a written assessment. The candidate viewed photo slides of fire scenes and then wrote down what he thought was the best course of action, including "his projected physical actions, his supervision of the firefighters under his command and his ability to deal with the dangers of the situation." But the Eighth Circuit correctly noted that the test required essentially "the written exercise of verbal skills" that were not necessarily important to the "physical, hands-on job" of fire captain, a job that involves "complex behaviors, good interpersonal skills, the ability to make decisions under tremendous pressure," and a host of other abilities.

Even if the substance of the test remains faithful to the job tasks, most test settings are quite unlike most job settings. In physical and psychological terms, test situations differ from actual work conditions. The mere possibility of test anxiety or evaluation apprehension may introduce variables into the test that would be unimportant on the job. In addition, the on-the-job presence of supervision, salary incentives, co-workers, and much longer periods of time in which to acclimate to job tasks renders testing and working conditions noncomparable. None of these transformations—in the nature of the job task, the setting in which it is conducted, or any others that might be introduced—are necessarily fatal to the job relatedness of the test. But these effects cannot be estimated with any degree of precision; they must be examined empirically.

199. Firefighter's Institute v. City of St. Louis, 21 Fair Empl. Prac. Cas. (BNA) 1140 (8th Cir. 1980).
200. Id. at 1147.
201. Id.
202. Id. at 1145.
203. See, e.g., supra note 135 and accompanying text.
205. The basic problem is that test performance may be determined by irrelevant factors that have nothing to do with job performance. Guion certainly recognizes this and suggests that these irrelevant sources of variance in test performance "should be checked out—not ignored." Guion, supra note 195, at 211 (emphasis in original). Yet there is only one reliable method with which such possibilities can be checked out—empirical studies that evaluate the employment test hypothesis.
The vagaries and potential dangers of content validation are clearly illustrated in National Education Association.206 In 1976, at the apparent urging of the test publisher (Educational Testing Service), the state of South Carolina authorized a validation study for the National Teacher Examinations that it had been using for the certification and compensation of its teachers.207 (Both the state and ETS apparently had cooperated for over twenty years in the use of the examinations without any such validation.208) The content validation proceeded in the following manner. Instructors from the state's twenty-five teachers colleges were asked to review the examination question “to decide whether each question on the tests involved subject matter that was a part of the curriculum at his or her teacher training institution,"209 and to estimate the overall relationship of the test to the state's school curriculum.210 Only test items which a majority of these reviewers judged “content appropriate” were used in establishing minimum passing scores on the test.211

Of the numerous problems with this “novel” design, several deserve emphasis. First, note that this was clearly not a content validation in which the job domain of teaching was sampled in any way. Rather, the procedure assessed knowledge of areas that were hypothesized to be important to teaching. But this hypothesis was never checked. Virtually no part of the job of teaching requires teachers to sit down under timed conditions and match wits with the authors of multiple choice questions. Moreover, the test failed to sample in any way many skills and talents that very likely are necessary for effective teaching: the ability to communicate what one knows, the ability to transcend one's college curriculum and make classroom material new and interesting, and a myriad of other interpersonal skills untouched by the test. It should be apparent that scores on these examinations bear no necessary relationship to performance as a teacher.212 Yet the state of South Carolina paid its teachers' salaries as a function of their scores on these tests, and it paid its black teachers less than whites as a func-

207. 445 F. Supp. at 1097-98.
208. Id. at 1102-03.
209. Id. at 1112.
210. Id.
211. Id. at 1112-13.
212. Indeed, existing research suggests that scores on the National Teacher Examinations correlate with very little, except perhaps undergraduate grade point average. For example, one extensive review concluded that “[g]iven the low correlations between National Teacher Examinations and ratings of on-the-job performance by principals and supervisors, it is difficult to justify the use of fixed cutoff [scores] in considering salary raises of teachers . . . contract assignment by school districts . . . provisional teaching certificates . . . and for differential rating on teaching certificates.” Quirk, Witten & Weinberg, Review of Studies of the Concurrent and Predictive Validity of the National Teacher Examinations, 43 Rev. Educ. Research 89, 109 (1973).
tion of the same tests. And despite its disclaimer that the tests "do not measure teaching skills," the district court used the scores to make inferences about teacher competence. None of this is justified by the method of content validation.

The lesson of National Education Association is not just that content validation alone cannot be used to evaluate the employment test hypothesis. The case speaks as well to the absence of clear and meaningful standards for the process of content validation itself. Viewed in the narrowest terms, the procedure used in this case could at most have indicated whether there was some overlap between the exam questions and some of the material taught in the state college curricula. No attempt appears to have been made in the content validation to determine whether the test tapped the most important material, material that teachers needed to know, or material that would make those who knew it better teachers. This narrow view of the test—that it measured only the retention of some material that had been taught in the state college curriculum—makes the job relatedness of the test and the state's rational interest in using it to determine pay differentials even more doubtful.

In addition, the content sampling was based on only a very small percentage of instructors—an average of less than one instructor from each college for each subject. Strictly speaking, of course, there was no content sampling in this case. Instead, there was post hoc evaluation by the college instructors of material already in the test. There is also no reason to believe these "experts" were particularly knowledgeable about classes other than the ones they themselves taught, which may or may not have been representative of the overwhelming majority of unsampled classes. One might also wonder whether college instructors are necessarily the best judges of what is taught in their classes. Moreover, if curricula varied systematically between colleges, and if those variations were related to the racial composition of the schools, the test could be racially biased and still "content valid" under this...

---

213. South Carolina's teacher pay scales are divided into four basic categories or levels. At the time of this litigation, 98% of the white teachers were in the two highest pay categories, and 51% of the black teachers were in the two lowest categories. 445 F. Supp. at 1116. For the long history of this racial disproportion, see supra note 123.


215. Id. at 1116.

216. To determine content validity, panels of "about 10 participants" were assigned to each of 18 separate "area examinations" (covering topics as broad as "English language and literature" and "social studies"), as well as to the separate sections of an overall "common examination." Since there are 25 state teachers colleges in South Carolina, far fewer than one instructor per campus could have participated in reviewing any content area. Id. at 1112. The ability of any one instructor to adequately represent this curriculum in these broad content areas, not only at his or her own campus but at other campuses as well, is questionable.
Curricula also change over time, so that items that were content valid at the time the study was done would not necessarily be valid for persons who had taken the test at an earlier time. If the test were re-administered to older teachers, they would be held accountable for material that perhaps they had never been taught. Moreover, substantial doubts were raised at trial about the degree to which at least some of the instructors followed the directions given for the assessment. Thus, there were numerous reasons to question the study as an adequate measure of even the narrowest form of content validity, let alone as a measure of job relatedness.

To get a better idea of the vagaries of content validation, compare the validation procedure outlined above with this passage from the APA Standards:

A school system was faced with the necessity of reducing its faculty. Reductions in force, according to policy, were to be based on teacher competence. However, decisions were in fact based on scores on a test that had been developed to evaluate the educational backgrounds of new teachers-college graduates. Nothing about the test established its validity as a measure of classroom effectiveness, nor was any local research conducted on this point. Its choice was, therefore, inappropriate.

Despite the similarity between this example of bad test use and the facts of National Education Association, and the numerous methodological questions that could be raised about the content validation the state had conducted, the principal author of these same APA Standards testified "in an unqualified fashion" at the National Educational Association trial that "the ETS study design met all the requirements of the APA Standards, the Division 14 Principles, and the EEOC Guidelines."220

The real danger posed by increased use of content validation procedures is that methods of empirical test validation, forced somewhat unwillingly on test psychologists by the litigation of the last decade, will be abandoned and mere opinion inserted in their place.221 What is

217. Indeed, there is evidence that the test is racially biased. See Medley & Quirk, *The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teacher Examinations*, 11 J. Educ. Measurement 235 (1974), who conclude that "whatever else this test may measure it has the potential of measuring a candidate's racial background with considerable accuracy," Id. at 244.
218. Id. at 1113-14.
220. 445 F. Supp. at 1113 (footnotes omitted).
221. The Division of Industrial-Organizational Psychology's Principles for the Validation and Use of Personnel Selection Procedure (1975) describe the elements of adequate content validation: (1) the job domain must be defined by the test developer; (2) the definition should be in terms of tasks, activities, or responsibilities, and possibly job knowledge; (3) the sample that is used in the test should cover important aspects of the job domain; and (4) the qualifications of the people who
at issue here is not just expense and convenience, but the very viability of many of these testing programs. When litigation forced test psychologists to "put the question to nature" by empirically examining the job relatedness of their tests, nature was not particularly kind to them. One psychologist commented early in the process that "we must reiterate the awful truth that nonsignificant validities against performance criteria are the general rule." This opinion was echoed by others:

We have been using a crude form of content validity to select experimental tests for empirical validation for many years. And I don't think anyone would argue with the observation that more of those experimental tests fall by the wayside than demonstrate significant validities.

Unlike empirical validation, it is almost always possible to construct a content valid test, whether or not the test is truly job-related. Indeed, the method of construction defines the validity. There is no empirical check against what is, at best, a plausible but completely untested hypothesis. In the extreme case, this lack of empirical checks may lead to what is termed "rational estimation" in which industrial psychologists simply make their best guess about the employment test hypothesis. The only published study to examine the ability of "highly knowledgeable" test psychologists to accurately predict actual relationships between test scores and job performance found that only one of every ten psychologists was able to do so at better than chance levels.

Note that each of these features is purely a matter of subjective judgment or opinion, not subject to clear and precise evaluation or to empirical disconfirmation.

223. Prien, supra note 165, at 173.
224. Prien has observed that in content validation, "the formal evidence is the 'paper trial' as compared to a statistical analysis." Id. at 167. But there is no professional consensus about what constitutes an adequate "paper trial," or how a test could fail such a trial. The APA Standards indicate that "content validity is determined by a set of operations, and one evaluates content validity by the thoroughness and care with which these operations have been conducted." 1974 APA Standards, supra note 219, at 29, but they do not indicate how "thoroughness" and "care" shall be defined. Indeed, the Foreword to the 1974 APA Standards expressed concern over the adequacy of the treatment given content validation, and promised a "companion volume" on the topic. None has appeared. Guion has reported that the members of the committee that worked on drafting the Standards "did not have clear ideas about the fundamental nature of what they meant by content validity or of its application to the field of employment testing." Guion, A Note on Taylor's "EEOC Guidelines on Content Validity," supra note 196, at 16.
225. Parry, Ability of Psychologists to Estimate Validities of Personnel Tests, 21 Personnel Psychology 139 (1968). This study is a marvelous example of the way in which numbers mean various things as a function of who reads them. In addition to the fact that only one psychologist in 10 in this study was able to predict at better than chance levels, correlations between psychologists' estimates of the test-job performance relationship and the actual relationship were all non-significant. In no instance was the overall estimate by the psychologists either negative or zero (i.e., they never predicted that there would be absolutely no relationship between test and job performance, or that high test scores would be associated with low job performance), yet in 10 of 42 instances this is precisely what occurred. In spite of all this, the author concluded that "the estimates of these psychologists are probably as good as the actual obtained coefficients." Id. at
Despite this fact, some test psychologists have touted a great future for rational estimation in employment testing "[s]ince criterion-related validity studies will frequently, perhaps typically, be technically infeasible, and in addition will sometimes be prohibitively costly . . ."\textsuperscript{226}

The logic of "rational estimation"—that opinion can substitute for empirical evaluation—is quite consistent with the logic of content validation. Predicting the validity of an entire test is not qualitatively different from estimating the job relatedness of specific test items, and neither approach can be endorsed in the context of disparate impact testing programs. While one might argue that content validity is an appropriate first step in the construction of a predictively valid test,\textsuperscript{227} the substitution of this preliminary step for the final empirical evaluation cannot be justified. Extensive reliance on the methodology of content validation promises a return to the days when "professionally developed employment tests" were simply ones that industrial psychologists believed to be adequate. Numerous testing cases under Title VII in which job relatedness could not be demonstrated underscore the folly of approving tests merely on the basis of such endorsements.

IV. BEYOND COMMON MODE OR PRACTICE: FUNDAMENTAL CHALLENGES TO THE WISDOM OF EMPLOYMENT TESTING

Technologies like employment testing possess an internal logic of their own. Accepted by critics as well as advocates, this logic focuses debate upon a narrowly circumscribed set of issues dictated by the technological imperatives of the very practice in question. But there is always a larger context for these practices that can be used to examine the internal logic as well as the technology it generates.

In the preceding pages I have outlined some of the technical limi-

\textsuperscript{226} Schmidt & Hunter, The Future of Criterion-Related Validity, 33 PERSONNEL PSYCHOLOGY 41, 48 (1980). The authors advocate the "estimation of employment test validities by experienced personnel psychologists," in part because "the cost of obtaining rational estimates should typically be relatively modest." \textit{Id.} Despite the suggestion that "courts also seem ready to look more favorably on employer arguments of high cost or difficulty in defense of their failure or refusal to comply with EEOC Guidelines relating to Title VII," Portwood & Schmidt, \textit{supra} note 114, at 180, the use of employment tests that are not demonstrably job-related, however modest in cost, hardly seems cost-efficient.

\textsuperscript{227} Cronbach has suggested, for example, that the various kinds of validity are "inseparable" and that the "tester cannot choose one kind as relevant and bypass the others." Cronbach, \textit{supra} note 129, at 45. Content sampling helps insure that predictively valid tests fairly represent the entire job domain.
tations and imperfections in common employment test practices. My basic argument has been that test users must demonstrate the predictive validity of test instruments that have an adverse impact on minority hiring or promotion. Any less stringent requirement fails to provide the necessary and convincing evidence of the test's relationship to job performance. The employment test hypothesis—that test scores predict job performance—remains merely untested speculation in the absence of such a demonstration. Disparate impact testing that denies jobs and advancement to minority workers cannot be justified on the basis of such speculation, nor can it be justified in the presence of non-discriminatory alternatives. In recent years, however, more fundamental challenges to the wisdom of employment testing have evolved. These challenges reach beyond the technical debate over predictive validity and into the larger context in which employment testing operates. Several of them warrant consideration.

Specifically, I first want to examine the changed conception of human nature that has emerged recently from contemporary psychological research, a conception markedly at odds with the one on which test theory is founded. Secondly, I will examine a new perspective on the nature of contemporary work, challenging the assumption that especially sophisticated skills are needed to perform in increasingly technological occupations. Finally, I will discuss a recent critique of the widespread belief that equal opportunity in the workplace can only follow equal opportunity in education.

A. Behavioral Specificity and Test Psychology

Studies of the predictive validity of employment tests have long documented—and lamented—the lack of generality in test-job performance relationships. The literature is replete with findings that tests which may predict job behavior in one situation nevertheless fail to demonstrate predictive validity in another. Indeed, one test psychologist suggested that the inability to generalize test-job performance relationships from one setting to another would prevent employment testing from achieving the status of a "science."229

For the most part, however, the limitations to these predictive relationships have been attributed to technical imperfections in the tests themselves. Test psychologists have assumed that more carefully constructed and sophisticated tests would eventually provide accurate measures of the essential or central traits and skills that would permit


229. Guion, Recruiting, Selection, and Job Placement, supra note 196.
the valid prediction of behavior in various settings. Recent analyses, however, suggest that there may be significant limitations to the degree of predictive accuracy that is possible to attain in employment testing. So long as the underlying theory of psychological testing was regarded as basically valid, deviations from acceptable or useful levels of accuracy could be viewed as largely technical problems, matters to be corrected through continued refinement of the test instruments themselves. But in recent years the very theoretical foundation on which employment testing rests has been shaken.

Test theory depends on the central assumption that people are the causal locus of behavior. To predict behavior, one measures the relevant characteristics of the people who perform it. Thus, traditional test theory “views the test as a measuring instrument intended to assign accurate numerical values to some quantitative attribute of the individual.”\(^{230}\) As an essential corollary to this basic premise, test theory (and the traditional theory of personality on which it is based) also assumes relative stability and consistency in human traits and abilities.\(^{231}\) It is the presumed stability of these attributes that provides the rationale for their measurement; if attributes were inconsistent or unstable, there would be little value or utility in measuring them. Thus, behavior sampled in the assessment or test setting is assumed to provide an accurate indication of behavior in some other, later setting. In the case of employment testing, of course, that later setting is the job itself. The assumption of stability and consistency means that the setting or situation is viewed as relatively unimportant in changing or modifying the essential traits or dispositions that are measured by tests and are assumed to determine individual behavior.\(^{232}\)

---


231. For example, Thorton & Benson, Industrial Psychologists as Expert Witnesses: Role Conflicts in Fair Employment Litigation, LAB. L.J. 417, 426 (1980), write that a “basic premise, that of individual differences, in the training of industrial psychologists . . . holds that individuals possess relatively stable and enduring characteristics that influence their effectiveness in organizations.”

232. One theorist, for example, defines “ability” as “that characteristic of an individual which permits him to behave adaptively, i.e., to cause the same result even though from time to time the circumstances vary.” Baldwin, The Role of an “Ability” Construct in a Theory of Behavior, in TALENT & SOCIETY 197 (D. McClelland, A. Baldwin, V. Bronfenbrenner, & F. Strodtbeck eds. 1958).

This exclusive focus on people as the causal locus of behavior explains the lack of attention given structural and organizational variables by industrial psychologists. See, e.g., Porter & Lawler, Properties of Organization Structure in Relation to Job Attitudes and Job Behavior, 64 PSYCHOLOGICAL BULL. 23 (1965), for a discussion of this, and a summary of evidence on the effects of organizational variables in industrial settings. See also Wallace, supra note 43, at 414:

We have long recognized that relationships between predictors and criteria can be greatly affected (moderated, if you will) by an unrecognized number and variety of situational factors, but only a very few of us have made any effort to manipulate or examine these factors in a systematic way so as to throw more light on the basic nature of our prediction problems.
These basic behavioral assumptions, however, are premised on a model of human personality and behavior that is now recognized as largely outmoded. Psychologists no longer view behavior as exclusively or even primarily the product of individual traits or personal dispositions. Rather, there is a "developing consensus that a personality discipline which is based implicitly on attributed dispositions is inadequate." Far more emphasis is placed upon the nature of the situation or context in which the behavior occurs. This change in emphasis has arisen in part because of dissatisfactions with trait conceptualizations of personality, in part because of low correlations obtained between measures of personality traits and various validity criteria, and in part because of growing evidence that substantial proportions of the variance in behavior are accounted for by situational and environmental variables.

The limitations of the traditional trait or dispositional model of behavior were presented most clearly by Walter Mischel in a critical review of the personality assessment literature. Mischel began by examining the ability of tests and assessment techniques to accurately predict subsequent behavior. He not only concluded that existing techniques were surprisingly poor at behavioral prediction but went on to question the basic premise on which they were founded. "[T]he assumption of massive behavioral similarity across diverse situations no longer is tenable." Rather, Mischel found much evidence for the situational specificity of behavior: people are highly sensitive to changes in immediate situational conditions. Indeed, the amount of variation in behavior that can be attributed to a measurable trait or personality characteristic is usually quite low. Although correlation coefficients between test scores and behavior might reach statistical significance if sample sizes are large enough, test scores, as previously noted, typically account for no more than ten percent of the variance in behavior. Instead of show-
ing broad, enduring stabilities in behavior, the assessment literature actually provided Mischel with much evidence for great specificity in behavior. He concluded that human behavior is highly dependent upon specific situational conditions in which it occurs. The nature of the immediate environment or situation plays a much larger role in influencing behavior than hitherto suspected.

Mischel’s work synthesized and extended existing doubts and criticisms of traditional personality theory and catalyzed a reexamination of the field’s basic assumptions. This reexamination and the emerging model of behavior it has produced have profound implications for the logic of testing and psychological assessment. Even those who continue to defend the traditional model of personality concede that “[s]kepticism regarding the usefulness of personality assessment seems to have become part of the conventional wisdom of contemporary psychology.” One prominent test theorist observed that “the conventional science of personality is close to its limits,” in part because of the inadequacies in test methodology:

We have been misled by the apparent objectivity of our data protocols. There is perfect consensus on the location of marks on a rating scale or questionnaire. Unfortunately, investigators are not interested in studying checkmarks.

Another succinctly described what he termed the “test-trait fallacy.” It “begins with the assumption that test scores are trait measures. The second assumption is that trait measures are basic properties of the person. It easily follows that test scores reflect basic properties of the per-

in academic settings are poor predictors of behavior in nonacademic environments. See Holland & Richards, Academic and Nonacademic Accomplishment: Correlated or Uncorrelated?, 56 J. Educ. Psychology 165 (1965). See also supra note 175 and accompanying text.

239. I do not mean to imply that there is uniform agreement over the exact shape and dimensions of this model. See, e.g., L. Tyler, Individuality (1979); Bem & Allen, On Predicting Some of the People Some of the Time: The Search for Cross-Situational Consistencies in Behavior, 81 Psychological Rev. 506 (1974); Bowers, Situationism in Psychology: An Analysis and Critique, 80 Psychological Rev. 307 (1973); Golding, Flies in the Ointment: Methodological Problems in the Analysis of the Percentage of Variance Due to Persons and Situations, 82 Psychological Bull. 278 (1975); Hogan, DeSoto & Solano, Traits, Tests, and Personality Research, 32 Am. Psychologist 255 (1977). Few would dispute, however, that the conception of human personality and behavior employed by most psychologists has been considerably transformed in the last decade.

240. As one writer graphically put it: “Tests, then, are samples of behavior, parts that hope to stand for a whole, buckets let down into the great sea of human nature to drag up enough specimens so that life down there can be recorded systematically.” R. Wernick, They’ve Got Your Number 38 (1956). Mischel and others have suggested that the sea of human nature ebbs and flows far more often than had been suspected and that, as a result, our test buckets frequently come up empty.


243. Id. at 2.

son." Each step in the progression now appears flawed and fallacious. And the flaws are not only in the tests, but also in the conceptual model of behavior on which they are based.

Although debate over the basic assumptions of test theory is contemporary and ongoing, similar judgments about the limitations of employment testing were reached years ago in a "classic study" in industrial psychology. The findings of the "Hawthorne experiments"—the field's best known research—strongly support Mischel's conclusions. First begun as a limited study on the effects of lighting on worker productivity, this research lasted for several years and eventually touched upon a large range of worker motivation issues. Perhaps the most crucial finding was that job performance depended upon a host of variables other than the worker's measured abilities; "tests of dexterity and intelligence showed no relation between capacity to perform and actual performance." Specific conditions in the workplace played a far more significant role than did test scores. Among the most important variables were interpersonal relationships, group norms, and other social psychological factors that were difficult to measure or predict beforehand. Group standards, more than individual ability, determined productivity. Elton Mayo, one of the study's chief researchers, concluded: "The belief that the behavior of an individual within the factory can be predicted before employment upon the basis of laborious and minute examination by tests of his mechanical and other capacities is mainly, if not wholly mistaken." Interestingly, the basic finding of this "classic" study has had little impact on personnel psychology.

Of course, to the extent that the most important performance variables cannot be assessed beforehand through techniques of individual measurement, the logic of the entire employment testing enterprise is

244. Id.

245. Id

246. For the standard account of these studies, see F. Roethlisberger & W. Dickson, Management and the Worker (1939). See also H. Handsberger, Hawthorne Revisited (1958); E. Mayo, The Human Problems of an Industrial Civilization (1933); Henderson & Mayo, The Effects of Social Environment, 18 J. Indus. Hygiene & Toxicology 7 (1936).

247. F. Roethlisberger & W. Dickson, supra note 246, at 446.


249. The test industry has often been criticized for failing to modify basic assumptions in response to research results. E.g., Fiske, supra note 233, at 3; Wallace, supra note 43, at 411. Compare Anastasi, Psychology, Psychologists, and Psychological Testing, 22 Am. Psychologist 297, 300 (1967):

Although the very essence of psychological testing is the measurement of behavior, testing today is not adequately assimilating relevant developments from the science of behavior. The refinements of test construction have far outstripped the tester's understanding of the behavior the tests are designed to measure.

A separate but related criticism has also been applied. Guthrie suggested that testing is "highly useful and practical work, but it has not contributed to psychological theory." Guthrie, Psychological Facts and Psychological Theory, 43 Psychological Bull. 1, 5 (1946). Lent, Aurbach & Leven, supra note 157, concede that this characterization "holds true" 25 years later. Id. at 520.
placed in question. If behavior is viewed more as the product of situations and less as a function of individual traits or dispositions, personnel selection that attempts to predict job performance by measuring the personal characteristics of job applicants will have questionable utility. In addition to casting doubt upon the logic of trait measurement, this changed conception of human nature carries three additional implications for employment testing. First, the importance of situational factors means that the conditions that prevail during test administration are potentially critical determinants of test performance. More careful attention must be paid to possible variations in the conditions that exist at the time tests are given, particularly since those variations may relate to racial status. Secondly, this new perspective suggests that assessment techniques must sample behavior from a setting or task as similar to actual work conditions as possible. The more the test differs in situational character from the actual job conditions, the less likely performance on the test will correlate with behavior on the job. "Work samples" or probationary work periods are far more likely to predict job performance than any attempt to sample broad underlying traits or skills. In fact, use of probationary work periods may be the only method accurately to assess performance potential. Finally, this perspective suggests that improvements in job performance will be attained more readily through changes in working conditions or situations than by changes in the characteristics of workers (or by changing the composition of the workforce through personnel selection).

Nonetheless, what once were the most widely shared assumptions in psychology—that people are the causal locus of behavior and that the measurement of individual trait differences can predict important differences in behavior—are now in doubt. This is the theoretical foundation of employment testing and in the last several years it has become extremely shaky. Although the controversy appears far from any definitive resolution, both sides seem to agree that previously unqualified assumptions about generalized traits and abilities are unwarranted, that behavior is far more sensitive to situational conditions than was earlier believed, and that the utility of any assessment technique can only be proven through careful empirical demonstration in the specific situation in which it is to be used. The implications of this changed conception of human behavior should not be lost on employment testing law, which seems to be on the verge of requiring even less scrutiny.

250. See supra note 135 and accompanying text.

251. For a brief review of studies in which "traits" are viewed as dependent variables that are altered by experimental treatments, see Anastasi, supra note 145, at 906-07. See also studies cited supra note 204; Porter & Lawler, supra note 232.
of test instruments at precisely the moment current psychological thinking suggests much more scrutiny may well be needed.

B. Testing and Job Complexity: Rethinking "Business Necessity"

Challenges to the use of employment tests are often rebutted by employers who claim that increases in the technological complexity of the workplace and the sophistication of tasks that workers are now required to perform render employment testing not only desirable, but necessary. Indeed, the concept of "business necessity" was given formal legal currency in Griggs, where the Supreme Court termed it the "touchstone" of permissible employment practices.252 In employment testing cases, attempts are made to harmonize the racial consequences of employment tests with the business needs of the employer.253

In this context, "business necessity" is a function of how well the test identifies certain applicant characteristics and of how much the employer needs workers who have those characteristics. Employers have generally argued that the complexity of the modern workplace means that workers must now be much more educated and more capable than in the past. Tests supposedly aid in this determination.254 The available evidence, however, suggests that many jobs have become significantly less complex and demanding, so much so that the more educated the workers, the more dissatisfied they are with their jobs.

For example, Ivar Berg and his colleagues have suggested that educational requirements are too heavily emphasized in hiring and promotion.255 Many workers, Berg reports, are simply "overeducated" for

252. 401 U.S. at 431.
253. Both the job-relatedness standard and the requirement that employers use available alternatives that are less discriminatory than disparate impact testing should be understood in this context. Tests that are not job-related perforce advance no legitimate business interest. Neither do tests that can be replaced by less discriminatory methods of employee selection.
254. The National Academy of Sciences offers this view: "With technological advance, the occupational mix of the labor force has come to include a greater proportion of technically trained workers.... These changes in skill requirements have an obvious implication for education: more and more specialized training is required. These training requirements also put greater demands on testing: more tests and more sophisticated tests are needed to determine training eligibility, to establish competence, and to evaluate the training organization." COMMITTEE ON ABILITY TESTING, ABILITY TESTING: USES, CONSEQUENCES, CONTROVERSIES, PART I, at 230-31 (1982). Of course, an approach that begins by assuming the necessity of testing or educational credentials is likely to overstate employers' needs. Consider this description of the rationale behind the Albemarle Paper Company's increased job qualifications:
In the 1950's the company began to modernize and install the technologically sophisticated machinery and equipment necessary to increase efficiency and productive capacity. Owing to this modernization, the company developed a strong need for skilled employees. In an effort to fill this need and to create a pool of employees who had the potential to progress to the jobs requiring more skill, Albemarle began to require that all new applicants for skilled lines of progression possess a high school diploma.

Johnson, supra note 72, at 1250. Yet the company was unable to demonstrate the relationship of this educational criterion to the successful performance of any of the jobs in question.
255. I. BERG, EDUCATION AND JOBS: THE GREAT TRAINING ROBBERY (1971). For a more
the demands of the work they actually perform. He cites several re-
search studies demonstrating that “Americans of diverse educational
achievements perform productive functions adequately and perhaps
well in all but a few professional occupations.”256 Berg also notes that
“variations in the characteristics of people performing adequately
within occupational groups have been found to be as great as variations
among these groups.”257 In their research, Berg and his colleagues
found that education and performance, as measured by absenteeism,
turnover, and productivity, were inversely related among both blue-col-
lar and white-collar workers. In one study of 125 branch offices of a
major metropolitan bank, Berg found the worst performance in offices
with workers whose educational achievements were the highest. Per-
formance was also poor for tellers in branch offices whose managers
stressed education in consultations with employees concerning their fu-
tures with the company. Berg also reported a number of instances in
which supervisors automatically assumed that their best workers were
their most educated workers when, in fact, the opposite was usually
ture.258

Thus, the increased emphasis on educational requirements cannot
be explained in terms of the increased demands of work. Rather, this
emphasis seems to be the product of an increasingly educated
workforce, coupled with the erroneous belief on the part of employers
that more education makes better workers. One regional director of the
Bureau of Labor Statistics described the phenomenon this way:

The completion of a high school education has become an important
requirement for entry into the labor market of today. Employers, find-
ing persons with high school diplomas becoming more available in a
period of rising educational attainment, have come to use the diploma
as a screening device, often seeking people with higher levels of educa-
tion even when job content is not necessarily becoming more complex
or requiring higher levels of skill. This has been true in many of the
rapidly growing job categories in the clerical, sales, and service

256. Id. at 41 (citing L. Thomas, The Occupational Structure and Education (1956)).
See M. Dunnette, Personnel Selection and Placement (1966); Thorndike, The Prediction of
Vocational Success, in Assessing Behavior: Readings in Educational and Psychological
Measurement (J. Flynn & H. Garner eds. 1967).

257. Id. (emphasis in original).

258. Id. at 93-94. In one company,
the less educated technicians received higher evaluations from supervisors and had
longer service than technicians with higher educational achievements in comparable
jobs; the managers, however, assumed that these “better” employees had completed
more years of schooling!

Id. at 16-17.
The same process probably occurs with employment tests; there is the tendency to hire and promote those persons who score highest on these tests, in spite of the lack of evidence to prove that the tests predict work performance or measure skills actually required for the job.

The contention that increased job complexity has made employment testing a vital business necessity is further challenged by Harry Braverman's research on the "degradation of work" in the twentieth century. In a critical analysis of the nature of contemporary work, Braverman makes a convincing argument that job tasks have become routinized and simplified by technology, rather than rendered more complex. Largely as a result of "scientific management," work operations have been separated, broken into their simplest parts, and assigned to different workers. Much of the labor force is engaged as "detail workers" who perform dissociated but repetitive tasks. This fragmentation of work process appears widespread in white-collar as well as blue-collar occupations.

Braverman's study also suggests that the apparent increase in the skills ratings for most jobs is an artifact of the classification process, rather than the product of a careful behavioral analysis of work. Standard classifications of job tasks automatically categorize workers as "semi-skilled" if their work is connected with machinery. Thus, as industry became more mechanized during the twentieth century, the

260. H. BRAVERMAN, LABOR AND MONOPOLY CAPITAL: THE DEGRADATION OF WORK IN THE TWENTIETH CENTURY (1974). See also J. BRIGHT, AUTOMATION AND MANAGEMENT (1958); Bright, The Relationship of Increasing Automation and Skill Requirements, in TECHNOLOGY AND THE AMERICAN ECONOMY: THE EMPLOYMENT IMPACT OF TECHNOLOGICAL CHANGE 220 (National Commission on Technology, Automation, and Economic Progress 1966), where the author states: "I suggest that excessive educational and skill specification is a serious mistake and potential hazard to our economic and social system. We will hurt individuals, raise labor costs improperly, create disillusion and resentment, and destroy valid job standards by setting standards that are not truly needed for a given task."
261. H. BRAVERMAN, supra note 260, at 70-83.
262. Id. at 81.
263. Id. at 72-73.
264. SPECIAL TASK FORCE TO THE SECRETARY OF HEW, WORK IN AMERICA, at xvi-xvii (1973). For example: The auto industry is the locus classicus of dissatisfying work; the assembly-line, its quintessential embodiment. But what is striking is the extent to which the dissatisfaction of the assemblyline and blue-collar worker is mirrored in white collar and even managerial positions. The office today, where work is segmented and authoritarian, is often a factory. For a growing number of jobs, there is little to distinguish them but the color of the worker's collar: computer keypunch operations and typing pools share much in common with the automobile assembly line.
265. H. BRAVERMAN, supra note 260, at 426-43.
266. Id. at 429.
skill classifications of most jobs were "upgraded." But this upgrading has had no necessary connection to the human work skills actually required by the job.

One index of the complexity of job tasks is the amount of time required to learn them. Braverman quotes the Department of Labor's Occupational Outlook Handbook's description of semi-skilled work: "Semi-skilled workers ordinarily receive only brief on-the-job training. Usually they are told exactly what to do and how to do it, and their work is supervised closely. They often repeat the same motions or the same jobs throughout the working day." According to Yale's Technology Project, for example, the average time cycle for assembly line work in the automobile industry was three minutes; a few hours, or a week, often sufficed to learn the job itself. These brief "training" periods are not limited to factory work. According to one writer, for example, a detailed manpower survey by the New York State Department of Labor showed that "approximately two-thirds of all the jobs in existence in that state involve such simple skills that they can be—and are—learned in a few days, weeks, or at most months of on-the-job training."

This general analysis of the nature of contemporary work must be viewed within the context of employment test practices that often assume, rather than demonstrate, the utility of the tests. One industrial psychologist characterized business executives as "amazingly gullible" because they "often purchase expensive 'employee selection' programs with no scientific evidence that the service offered has any value whatever." A few years later, another candidly observed that "we and our sponsors continue to 'give [tests] a whirl' " with no good evidence that they improve the quality of the workforce. One study showed that seventy percent of the firms interviewed had adopted tests without the aid of job analysis or professional consultation, and that less than one company in five had any evidence that the tests improved their employee selection. And in one recent case, a city purchased

267. Id. at 424.
tests from a consulting firm that merely "told the city that all of the tests sold to it were validated," without giving any indication of how or to what extent they were validated. 274

Undoubtedly guided by the general belief that technical job tasks require an increasingly skilled workforce, many employers conclude that testing is necessary for effective personnel selection. This belief seems to obviate the need for convincing evidence of the utility of testing. However, the work of Berg and Braverman, demonstrates that technical advances in the workplace have not drastically increased job complexity. The business necessity of employment tests declines as a result. As a society, we can and should deplore the routinization of work. More importantly, we should not add the insult of discriminatory employment testing to the injury of routine work by disqualifying people from jobs because they lack skills that are not actually required for the work in question. Fair employment practices should go hand in hand with the upgrading of jobs. Illusory levels of job skills, however, should not be employed as a legal fiction to maintain claims of business necessity where none, in fact, exist.

C. Choosing Equality: Schools or Jobs?

The practice of employment testing assumes that differences in test scores reflect real differences in ability, such that those who score lower have less ability. It also assumes that those ability deficits are relatively stable properties of the individuals themselves and are not easily correctable on the job. One test theorist writes that:

To criticize tests because they reveal cultural influences is to miss the essential nature of tests. . . . The same cultural differentials that impair an individual's test performance are likely to handicap him in school work, job performance, or whatever other subsequent achievement we are trying to predict. 275

Few employers openly assert that the ability or aptitude differences supposedly revealed by tests are "innate." Instead, the more

---


275. Anastasi, supra note 249, at 299.
common assumption is that the differences can be attributed to early deprivations, to negative past experiences, and, especially, to inferior education. Minorities fare less favorably in the marketplace, the argument runs, because they are less skilled as a result of their poor education. Many defenders of employment testing go so far as to claim that any restrictions on the use of ability tests or lowering of the minimum scores that are permitted would actually do minorities a disservice by easing the pressure to develop better minority education programs. The place to eradicate the effects of past discrimination, test advocates argue, is in the schools, not in the workplace.276

There is a painful irony to proposals that locate the solution to the disparate impact of employment testing solely in better educational opportunities for minority groups. Such proposals ignore the disparate impact that standardized tests have on minorities in schools, where tests are given and relied upon even more heavily than in employment settings.277 One of the criteria the Supreme Court has legitimated for the distribution of educational opportunities is “ability,”278 and in educational settings ability is invariably indexed by performance on standardized tests.279 These tests are not only used as admissions criteria by elite college and graduate programs that enroll proportionately few minorities, but have been employed to “track” a disproportionate number of minority students into inferior programs within public schools as well.280 Even when students are “mainstreamed” into the same classes, test scores can become self-fulfilling prophecies by creating differential expectations in teachers that alter both teaching and learning.281 Moreover, standardized tests are used to channel students into occupational and career areas, as well as to shape their self-con-

276. See, e.g., Lerner supra, note 112, at 311-16.
277. In one study, for example, 75 of 75 secondary school principals and 713 of 714 elementary school principals questioned reported using standardized tests with their pupils. Goslin, supra note 1, at 13-19.
279. See A. NAIRN, supra note 176, at 28-54, for an analysis of the influence that the testing establishment exercises in educational decision-making.
281. E.g., R. ROSENTHAL & L. JACOBSON, PYGMAIION IN THE CLASSROOM: TEACHER EXPECTATION AND PUPILS’ INTELLIGENT DEVELOPMENT (1968). See also D. GOSLIN, supra note 1, at 52. Goslin’s study showed that teachers are convinced of the validity of standardized tests, most of which are no more impressively validated than the employment tests I have been discussing. Over 38 percent of the teachers and 56 percent of the guidance counselors surveyed felt intelligence or scholastic aptitude test scores were “the single most accurate measure of a student’s intellectual ability.”
concepts and aspirations. Nonetheless, ability and aptitude testing in schools is not afforded the same degree of legal scrutiny that has been applied to employment tests.

Thus, to absolve employment testing of the responsibility for racial inequalities in employment opportunity, by focusing instead on education, is merely to shift the point at which "tests" make a difference. Aside from the shortsightedness of these proposals, there is another reason to challenge them. Anthropologist John Ogbu has presented a forceful argument that the assumed relationship between poor education and poor employment may well be reversed:

[B]lacks and similarly placed minority groups often reject academic competition with members of the dominant groups. The reason they fail to work hard in school seems to be, in part, that such efforts have not traditionally benefited members of their group. In terms of ability and training, they have generally received lower social and occupational rewards.

Rather than poor education causing poor work performance and lower wages, Ogbu suggests it is the disparity in job success and compensation that causes poor school performance. Blacks know that superior performance in school will not lead to occupational success and are influenced in school by the reality of a "job ceiling"—the lowered level of occupational success they may reasonably anticipate achieving. Children's achievement motivation is undoubtedly affected by what parents, teachers, and the media communicate to them about opportunities and social mobility. Behavior is influenced not only by its immediate consequences, but by anticipated rewards and expectations.

Blacks learn that they have been excluded traditionally from occupa-

---

282. Gosin found that the two most important functions of standardized test scores, as reported by high school administrators, were to "help pupils gain a better understanding of their strengths and weaknesses" (rated by 71.6% as a "very important" use) and to "help in educational and vocational counseling of pupils" (69.3%). D. Goslin, supra note 1, at 20. Although given in virtually all schools, standardized tests "are used most extensively in schools having a heterogeneous student body (both college and noncollege-bound students in which problems of pupil guidance classification and evaluation are likely to be more acute." Id. at 25 (emphasis in original).


283. See, e.g., White, Culturally Biased Testing and Predictive Invalidity: Putting Them on the Record, 14 Harv. C.R.-C.L. L. Rev. 89 (1979). But see Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979), where a federal district court did scrutinize intelligence tests used to track students into EMR classes, and Bersoff, Regarding Psychologists Testify: Legal Regulation of Psychological Assessment in the Public Schools, 39 Md. L. Rev. 27 (1979), who suggests that psychological tests are now legally overregulated in public schools.

284. J. Ogbu, supra note 282, at 4 passim.

285. Id. at 4.

286. A substantial amount of human learning is vicarious, observational, and anticipatory in nature. See A. Bandura, PRINCIPLES OF BEHAVIOR MODIFICATION (1969), and studies cited therein, for a review. See also R. LeVine, DREAMS AND NEEDS: ACHIEVEMENT MOTIVATION IN NIGERIA (1967).
tions above the job ceiling. This knowledge undermines their academic motivation:

[M]inorities perceive their future chances for jobs and other benefits of education as limited, [so] they are not so strongly motivated as the dominant group members to persevere in their schoolwork. ... [Secondly,] cognitive, linguistic, motivational, and other skills that dominant-group children take with them to school are intimately related to the types of such skills required and promoted by the social and occupational roles they will play as adult members of society. 287

Thus Ogbu would reverse the standard “blacks get poorer jobs because they are more poorly educated” explanation and concentrate instead on the effect job discrimination has on educational achievement. For example, he cites data showing that the “educational gap” between blacks and whites in America is much narrower than the “occupational gap.” One study shows that blacks with some college training work in service and laborer occupations “in numbers five times greater than whites with similar training,” and that ten percent of black women who finish college end up as domestic workers. 288 Differences between the unemployment rates of blacks and whites are greatest among those with four or more years of college. 289 Moreover, the educational gap is closing more rapidly than its occupational counterpart. Ogbu concludes: “These discrepancies in the relationship between education and occupational and employment status for the two racial castes lead one to conclude that blacks and whites are not selected for their occupations on the basis of the same objective criteria.” 290

Contrary to the arguments of employers and test advocates, Ogbu’s analysis suggests that improved educational opportunity—while desirable in its own right—will not automatically lead to greater success in employment. In fact, equal occupational opportunity may well be a prerequisite to educational achievement among minorities. But the practice of employment testing helps to maintain the disparity between minority employment and education gaps. 291 Continued use of employment tests that have a disparate impact on minorities will not only perpetuate occupational inequities, but impede the development of equal education as well.

287. J. OGBU, supra note 282, at 41.
288. P. NORgren & S. HILL, TowARD FAIR EMPLOYMENT (1964); kahn, The Economics of Inequality, in Negroes and Jobs: A Book of Readings (L. Ferman, J. kornbluh, & J. Miller eds. 1968).
290. Id.
291. Most non-educational forms of “cultural deprivation” that might be used to account for racial disparities in employment performance can be ascribed directly to poverty. The solution to poverty entails, at least, equal employment opportunity.
V. STRICHER SCRUTINY: STRENGTHENING FAIR EMPLOYMENT TEST POLICIES

Employment testing has become an important industry in this country. It is unrealistic to advocate the sudden termination of such widespread practices in the United States. Yet there exist fundamental challenges to the employment test industry that undermine both its legitimacy and its viability. If human behavior is regarded as highly situation-specific, the logic of traditional trait and aptitude assessment becomes tenuous. Standard test assessment must be seen as inherently limited, and new situation-centered models of evaluation must be substituted in its place. If technological changes in the workplace have actually reduced the complexity of many job tasks, elaborate and highly discriminating personnel selection cannot be viewed as a business necessity. Larger numbers of persons are potentially acceptable as employees, the costs of employment testing are less justified, and on-the-job training is more feasible. If minority performance is significantly hindered by the reality of unequal opportunity, continued allocation of jobs on the basis of stringent test performance criteria will retard the progress of racial justice. Indeed, equality of opportunity may be a precondition for equality of performance.

These challenges reach beyond debates over minor technical imperfections and the fine-tuning of test methodology. They argue for the stricter scrutiny of testing practices, especially ones that have predictably adverse consequences for constitutionally protected groups. In this context, several proposals can be made that help to clarify the role of employment testing in employment discrimination and to reduce its disparate impact on minorities.

A. Employment Tests and Decision Theory

In typical employment testing cases, employers appear to be acting quite irrationally. They insist upon using tests with little demonstrable predictive validity, and they justify this testing on grounds of efficiency and business necessity. But what efficiency attaches to the use of a predictively invalid test? This apparent irrationality is best understood if a different conceptualization of the testing process is used, one that recognizes the process in terms of decision theory rather than classical test validity.

Employment test measurement and prediction are simply techni-
cal components in a larger system designed to make decisions about employment. In practical terms, the number of correct decisions that an employment test produces is more important than the size of validity coefficients it generates. Decisionmakers in this context use test scores and other data to make decisions about assigning persons to "treatments"—to offer them employment or not. Information that is collected beforehand is used to construct a predictive index of likely job performance. But the index does not straightforwardly dictate the decision. Rather, some decision rule must be established to operationalize the decision itself. Although the reliability and validity of the testing devices are a part of this process, calculation of these coefficients does not end the decisionmaking process. As one psychologist observes: "Whereas this step has sometimes been considered the final contribution of [test] assessment, it, in fact, only sets the stage for the decision problem." What remains to be done is to assign values to the various possible outcomes of the decision. Not all outcomes are of equal importance, nor are the same outcomes valued identically by the different participants and interested parties in the decision. This stage of the process, often overlooked in classical test theory, is the most vexing. As Cronbach and Gleser put it: "The assignment of values to outcomes is the Achilles heel of decision theory."

The "correct" decision in an employment selection means hiring or promoting those persons who will perform at a certain level of proficiency and rejecting applicants who will not. An "incorrect" decision, of course, involves either hiring or promoting someone who will not perform at the requisite level (a "false positive"), or rejecting someone who will ("false negative"). Since even the best employment tests are far from perfect predictors of job performance, employers make incorrect decisions of both kinds. Depending upon how the decision rule is framed, however, more or less of one or the other error is likely to occur. The particular value assigned by the decisionmakers to each

294. A variety of factors, of which predictive validity is only one, influence the size of a test's contribution to a "correct" decision. That is, decision theory suggests that a number of other factors are important in addition to test validity. In an employment discrimination context, in which the test in question has already been shown to have a disparate impact on minorities, an employment test with unimpressive predictive validity coefficients might still marginally increase the number of "correct" decisions an employer could make, but do so at the expense of errors that are intolerable from the perspective of fair employment policy.

295. J. WIGGINS, supra note 293, at 225.

296. L. CRONBACH & G. GLESER, supra note 230, at 121.

297. This terminology is employed in signal detection research in which subjects are asked to identify the presence of a target stimulus or "signal." Subjects can err either by failing to perceive the signal when it is present (false negative) or reporting its presence when it is not (false positive). False positive errors are also referred to as "Type I" errors, false negatives as "Type II," in psychological parlance.
type of error will depend upon such factors as the prevailing job market, the size of the company's investment in each new employee, and the consequences of inadequate job performance. For example, the larger the prevailing relevant job market, the more tolerant an employer is likely to be of rejecting an applicant who would have performed at the required or desired level—a false negative.

Values assigned to these two different kinds of errors help determine the "cutoff" score, below which applicants are not hired or promoted. But employers are likely to adopt a relatively conservative decision rule that weighs heavily the number of false positives that occur, while discounting false negatives. For example, Cronbach and Glaser suggest that employers should disregard entirely the costs of false negatives in hiring:

In institutional selection with fixed quota where men rejected leave the institution, it is not meaningful to consider misses, however. Rejection of men of good quality does not decrease the output of the institution; their frequency bears only on "what might have been." 298 Indeed, this is the perspective that has dominated personnel decision-making. 299 Of course, it is the opposite of the individual applicant's perspective, in which the disutility of a false negative far outweighs that of a false positive. And, from the perspective of fair employment policy, it is unacceptably one-sided.

Yet this one-sidedness is built explicitly into employment test practices. "In academic and industrial selection, assessment is in the service of an institution rather than of the individual." 300 Test psychologists are trained explicitly in this perspective. For example, one well-known textbook instructs:

Given a choice between acts (placements) in the face of uncertain outcomes (performance), the effective decision maker behaves in such a fashion as to maximize the values (profits) of the institution he represents. . . . Assessment psychologists who drag their feet or are otherwise indifferent to the principle of maximization are implicitly supporting irrational behavior in areas of great social importance. 301

298. L. CRONBACH & G. GLESER, supra note 230, at 52.
299. Compare Wallace, supra note 43, who writes that "[s]omehow, we seem to have convinced our clients, and even ourselves, that the application of our selection techniques is, at worst, neutral in its effect." Id. at 411. Yet employment testing proceeds "in blissful oblivion of its true cost, i.e., the rejection of potentially adequate employees." Id. at 412. Compare also Fincher, supra note 73, at 495: "The cost of a poor choice may be obvious in an employee whose performance is unsatisfactory, but the possible performance of the unhired remains unknown." See also supra note 176 and accompanying text.
301. J. WIGGINS, supra note 293, at 224. See L. CRONBACH & G. GLESER, supra note 230, at 8: "One seeks the decision or policy which will yield as much benefit as possible to the institution. Individual interests may be taken into account, but only insofar as they affect the realization of the goals of the institution." See also D. GOSLIN, supra note 4, at 96:
Profit maximization, however, cannot be the sole determinant of a national policy of fair employment. By weighing exclusively the interests of the institution or employer, the test industry actually obstructs and subverts fair employment policy, a policy that must adjust business interest to competing social concerns.

In a sense, employment discrimination law attempts to harmonize the competing values in personnel decisions. In “restrictive” testing decisions, courts have told employers essentially that false negatives in personnel decisions deserve more serious attention, at least when the persons falsely rejected are minorities. A hiring or promotion system that has tolerated these kinds of mistakes in the past cannot continue to do so when the consequences of those mistakes fall disproportionately upon minority applicants. The employer’s interest in minimizing one kind of error must be balanced against a constitutional and statutory interest in equal treatment. Judicial intervention in this area can be seen less as a demand for higher validity coefficients in employment tests and more as a restructuring of the values that influence the decision process. “Personnel selection is a political act,” and as such must be subject to various legitimate political constraints. One consequence of these constraints is that the employer’s interest in minimizing hiring costs must be moderated against the social interest in maintaining a fair and just employment system. Thus, predictively valid tests, however uncommon and essential, are still not decisive. They should be evaluated in this larger context of personnel decisionmaking.

Explicit decision theory analyses facilitate this balancing in several ways. For one, the number of both types of potential hiring errors can be estimated beforehand. That is, by adjusting the decision rule, one can predict in advance how many errors of either type will occur. Courts could well require data on challenged employment tests to be calculated in terms of the number of minority persons in a given applicant pool who—on the basis of the tests that were used, validity coefficients for the particular group, cutoff points established, and so on—were incorrectly rejected for employment. Evaluation of tests in these terms conveys a more realistic picture of the consequences of low validity and disparate impact testing—the incorrect rejection of minority employment applicants. At the same time, one psychologist has used

Business and industrial testing is essentially one-sided. The test’s primary function is to benefit the company, not the individual. To the extent that being placed in a position that requires the use of one’s abilities is beneficial to the subject, one may gain from being tested, but the rationale for industrial testing is the company’s potential gain in productivity as a result of making maximum use of personnel resources.

302. Cronbach, supra note 129, at 49.

303. Recent suggestions that much larger sample sizes should be used in the calculation of predictive validity coefficients in order to obtain more statistically significant results, e.g., Schmidt & Hunter, supra note 226, at 43, overlook an important point. In the context of fair employment policy, less concern should be given to whether there is some measurable relationship between
decision theory to demonstrate that various models of "fairness" in hiring actually have little appreciable impact on the presumed "quality" of the workforce, despite their substantial impact on minority applicants.\(^{304}\)

Decision theory also permits quantification of values or utilities.\(^{305}\) It forces decisionmakers to make explicit the preferences that control their selection:

The unique feature of decision theory or utility theory is that it specifies evaluations by means of a payoff matrix or by conversion of the criterion to utility units. The values are thus plainly revealed and open to criticism. This is an asset rather than a defect of this system, as compared with systems where value judgments are imbedded and often pass unrecognized.\(^{306}\)

This feature may make the development of a uniform policy of fair employment testing practices more feasible. When values are made explicit they can be more easily scrutinized and regulated. At the same time, however, decision theory is not immune to the tendency to quantify only that which is most easily measured. Emphasis on economic costs and benefits to the exclusion of their human, social, and political counterparts should be avoided here as well.\(^{307}\)

**B. Imposing Standards: Beyond Professional Self-Validation**

In early attempts to evaluate employment tests, courts focused on the term "professionally developed," taken from the Civil Rights Act of 1964, as an operational definition of acceptability.\(^{308}\) Employment tests were acceptable if professional employment testers had developed them.\(^{309}\) But subsequent litigation found the professionals themselves in conflict over many of the tests they and their colleagues had developed. When the Supreme Court finally intervened in *Griggs* and *Albemarle*, it imposed a tentative set of standards, fashioned largely from EEOC guidelines that had held employment testers to the highest pro-
fessional standards. Yet *Washington v. Davis* and its ensuing commentary, along with possible shifts in federal agency policy, threaten a return to that earlier period of professional self-validation. This development is ill-advised.

While professional test developers and administrators may be capable of refining test instruments and calculating their predictive capacity, they are *not* ideally suited to decide upon the values that should be attached to each outcome in the employment decision process. They have their own interests at stake in these situations, and their testing practices are explicitly dependent upon the goals of their employers. In this context, there is no reason to expect interested professionals to be especially cognizant of or sensitive to the moral, social, and constitutional issues involved in the creation of a fair employment test policy.

The record of professional psychology on these issues is neither entirely laudable nor completely impeachable. I have already recounted psychology's early role in perpetuating the "brutally pessimistic" view that human traits and abilities were inborn and racially distributed. These beliefs were most certainly *not* "soon condemned by the leaders of the profession," as some have suggested. Psychology continued to provide scientific legitimation for the highly dispositional trait-centered view of behavior that too often was used "to categorize people enduringly into fixed slots," and continued "to assume that these slot positions were sufficiently informative to predict specific behavior and to make extensive decisions about a person's whole life." This is a view of behavior that is widely held by the general public, while it is increasingly regarded as outmoded by the profession.

When testing became "big business" for psychology, internal professional criticism of widespread abuses also developed. For the

---

310. See supra notes 300-01 and accompanying text.
311. See Part I supra.
312. Cronbach, supra note 129, at 39. Indeed, the article Cronbach cites in support of this statement, Freeman, *A Referendum of Psychologists*, 107 Century Mag. 237 (1923), actually provides clear evidence to the contrary. For example, all of the respondents to this "referendum" agreed with the statement that "marked differences in intellectual capacity exist within groups of fairly equal educational and environmental opportunity and that these differences are inherent." *Id.* at 242. But see Samelson, supra note 45.

*[I]t should be apparent that the use of psychological tests as selection devices far exceeds the availability of trained personnel competent to install and interpret them. The growth of 'do it yourself' selection testing by untrained or partially trained amateurs has undoubtedly been encouraged by the ease with which tests, particularly aptitude tests, have been made available to all who wish to purchase them. Even the most reputable of test publishers have lowered or removed the barriers that previously restricted the distribution particularly of aptitude and intelligence tests.*
most part, however, this criticism focused on the professional credentials of the test users or administrators (i.e., that they were not psychologists) and the insufficiently scientific, overly commercial manner in which tests were marketed and sold. These early statements reflected the tensions inherent in the attempt to conduct science as lucrative business. The profession was learning that profit motives did not mix well with the objective search for truth.

When widespread criticism, however, was later directed at test practices from sources outside the profession, it was often characterized by psychologists as objectionable,\textsuperscript{315} and "extravagant and unfair."\textsuperscript{316} In the wake of this criticism, one professional tester suggested that psychologists "might ponder whether the need is for further self-restriction (as opposed to renewed affirmation of existing principles) or for improved public relations," and he cautioned his colleagues to "be especially careful not to let it be inferred that any change in our standards for psychological assessment necessarily reflects a general admission of past guilt."\textsuperscript{317} Occasionally psychology turned its unique conceptual weapons against those who criticized its testing. In a discussion of the "basic sources" of test criticism, for example, one psychologist suggested that "[t]hose who object to psychological testing may do so by virtue of their personality characteristics."\textsuperscript{318} And an APA committee proposed that a study be conducted of "persons favorably and unfavorably disposed toward testing," to see whether, \textit{inter alia}, "people who oppose testing (are) like those who oppose fluoridation."\textsuperscript{319}

More recent criticism of testing from outside the profession has focused primarily upon its discriminatory impact. Unlike the earlier criticism, challenges to the racially discriminatory aspects of testing have resulted in laws, agency regulation, and seemingly endless litigation. Much of the professional reaction to this legal control has centered on the EEOC Guidelines, particularly as interpreted by the federal

\textsuperscript{316} Messick, \textit{supra} note 300, at 137.
\textsuperscript{317} Id.
\textsuperscript{318} Brim, \textit{American Attitudes Toward Intelligence Tests}, 20 Am. Psychologist 125, 128 (1965). Brim elaborated:

It is likely, for example, that strong opposition exists among those people who are distinctly hostile to any self-examination, introspection, or self-understanding. These are the people who are also authoritarian in interpersonal relations, intolerant of diversity in ideology or beliefs, and strongly opposed to most forms of social change. The association of this personality type with extremist groups in our country, especially the extreme right-wing political groups, is commonly recognized, and it is understandable that such groups are pushing for legislation banning the use of tests.

\textit{Id.} Before testing was criticized for its impact on minority groups, right-wing political groups objected to it on privacy grounds. In present times, the above "psychodiagnosis" of test critics would undergo considerable change.

courts. Test psychologists complained that the Guidelines were vague, overly stringent, unrealistic, and out-of-date.\footnote{320} Ironically, these very guidelines drew much of their conceptual framework from existing psychological standards. The public record alone makes it clear that psychologists and the APA Standards have been widely consulted in each revision of the EEOC Guidelines.\footnote{321} Psychologists, however, do not speak with a single voice on these issues.

The new Uniform Guidelines\footnote{322} appear to move even closer to both the letter and spirit of the APA Standards than any of the earlier EEOC versions. It is impossible to predict with complete certainty how the new Guidelines will be interpreted by the courts. But their apparent compatibility with the psychological Standards has been acknowledged by the APA, which noted the high degree of consistency between the two documents.\footnote{323} Despite the benefits that may come from terminating fruitless bickering between test psychologists and their legal regulators, a shift in EEOC policy in the direction of the psychological Standards should not be applauded as an unqualified blessing. Several serious problems may be posed by an excessive reliance on the APA Standards in employment testing cases and by the possibility that courts may use them exclusively to gauge the propriety of challenged test practices.

The first problem is that these standards were not designed for application solely in employment settings. Of necessity, then, they fail to reach and resolve important employment discrimination issues. At the same time, their general purpose affects the basic substance of the stan-

\footnote{320. E.g., Cronbach, supra note 129, at 37; Fincher, supra note 73, at 494; Lerner, supra note 112; Novick & Ellis, Equal Opportunity in Educational and Employment Selection, 32 AM. PSYCHOLOGIST 306 (1977). See also Hunt, Civil Service Testing and Affirmative Action: A Psychologist's Perspective, 44 U. CIN. L. REV. 690 (1975), who suggested that psychologists have not participated fully in the development of legal regulations; Wollack, Content Validity: Its Legal and Psychometric Basis, 5 PUB. PERSONNEL MGMT. 397 (1976), who argued that the courts have employed a "clearly erroneous" and overly restrictive interpretation of the EEOC Guidelines; and Muchinsky, Book Review, 29 PERSONNEL PSYCHOLOGY 491, 492 (1976), who lamented, in cross-cultural observation, that "the English are where we would like to be, relatively free of the stranglehold of personnel selection brought about by legal intervention."}

\footnote{321. The first set of EEOC Guidelines on Employment Testing Procedures were published with a statement of a panel of consulting psychologists appended. See EMPL. PRAC. GUIDE (CCH) ¶ 16,904 (1967). The revised version of these Guidelines makes explicit reference to the APA Standards. See 35 Fed. Reg. 12,333 (1970). See also Taylor, The EEOC Guidelines on Content Validity and Their Application to the Assessment Center Method, 1 J. ASSESSMENT CENTER TECH. 9 (1978): "The psychological staff of the Commission, however, have used the APA 1974 Standards as the reference source since its issuance, and are in complete accord with its treatment of test validation." But see Guion, A Note on Taylor's "EEOC Guidelines on Content Validity," supra note 196.}

\footnote{322. See supra notes 192-94.}

\footnote{323. 45 Fed. Reg. 29,530 (1980). See also Lerner, The Supreme Court and the APA, AERA, NCME Test Standards: Past References and Future Possibilities, 33 AM. PSYCHOLOGIST 915 (1978).}
EMPLOYMENT TESTS AND DISCRIMINATION

Because the Standards were designed to apply to a wide range of test uses—including research and experimental testing—they include provisions that are not necessarily appropriate in employment contexts where, for example, the consequences of test misuse are grave and often irrevocable.

Secondly, the APA Standards not surprisingly reflect an almost irrefutable presumption in favor of some sort of testing, even under circumstances that are recognized as questionable and undesirable:

Tests are often developed and used in circumstances that lead to maintaining less than the highest standards of technical excellence. We do not intend to discourage those who must make assessments of people from doing the best they can with whatever training and collaborative resources are available to them.

These are more than concessions to pragmatism. They reflect a confirmed view of the propriety of testing, despite the risk that under some circumstances tests will be improperly developed and used. The Standards fail to define any situations in which testing or assessment should not be done. They assert instead that if those responsible for making selection decisions do not use tests, "they will use less dependable methods of assessment." The possibility that some tests are too unreliable—or discriminatory—to be used at all, eludes them.

Finally, and most importantly, the Standards are not really standards at all, but are "statements of ideals or goals." The manual disclaims any intention to require test developers or users to follow them: "This document . . . is not written as law." The Standards

324. There is an obvious tension between experimentation and regulation. Although the 1974 Standards expressed "concern about problems like invasions of privacy and discrimination against members of groups such as minorities and women," id. at 1, they are concerned also about being "unduly rigid" since standards "should not bind the producer of a novel test to an inappropriate procedure." Id. at 6. Both the 1954 and 1966 versions of the APA Standards expressed concern that "issuing specifications for tests could indeed discourage the development of new types of tests." 1966 STANDARDS, supra note 84, at 2. In fact, the 1954 version quoted approvingly these social Darwinist-like sentiments: "Let very many tests be tried, each new investigator introducing his own modification; and then, the worthless will gradually be eliminated and the fittest will survive." APA TECHNICAL RECOMMENDATIONS FOR PSYCHOLOGICAL TESTS AND DIAGNOSTIC TECHNIQUES 1 (1954) [hereinafter cited as 1954 APA RECOMMENDATIONS]. The document did not discuss at whose expense this trial-and-error process would occur. In the desire to give psychologists the flexibility to experiment and innovate, these standards provided much latitude for questionable test practices that have had serious real world consequences.


326. Id. at 3 n.1. Test users are told that "[t]he standard is to do the best one can," id. at 71, but, apparently, always to do something.

327. Id. at 8. This caveat undermines any effective use of these standards as regulatory tools. Apparently, it is ethical for psychologists to aspire to reach these "ideals or goals," but not unethical for them to fail. The provisions of the Standards are grouped into three levels, as a function of their importance. Even provisions in the "essential" or most important level, however, allow that "when a test developer or test user fails to satisfy these requirements, he should do so only as a considered judgment." Id. at 6.

328. Id. at 8 (emphasis in original).
specifically refuse "to call for a particular level of validity or reliability, or otherwise to establish technical test specifications for specific tests." Further, because the Standards are ideals or goals rather than required specifications, "it would not be appropriate for test developers or test users to state that a manual or procedure 'satisfies' or 'follows' these standards." Largely because they were intended as goals rather than specifications they are unworkably vague and imprecise as guidelines for employment testing litigation. At crucial points, for example, it is simply impossible to determine what practices would constitute "violation" of the letter or spirit of various provisions. Since violations were not really contemplated by these Standards, of course, this is understandable. But it is no less problematic.

One psychologist has accurately characterized the APA's stance on test regulation: "Along with the Educational Testing Service the APA answered coherently and constructively the arguments of test critics." It "answered" but did not confirm, extend, or in any way lead this criticism. Moreover, as a professional guild designed to protect and advance the interests of its members, the APA cannot be expected to formulate and impose rigorous standards of test regulation. Such standards must be externally created, as well as enforced, by an independent agency.

Employers are similarly ill-suited to regulate their own testing. Use of employment tests by personnel managers is often based on little or no valid information. One study concluded that "there was essentially no systematic relationship between [employer] perception of ef-

329. Id. at 5.
331. It should be noted that the APA did not have to issue these standards, and it is a tribute to the ethical concerns of the organization that it did. At the same time, however, these standards should not be confused with regulations. They began as, and will continue to be, a series of suggestions to test developers and publishers about the kind of information that "is most helpful to the test consumer," 1954 APA RECOMMENDATIONS, supra note 324, at 1, and therefore should be included in test manuals. The 1974 version also contains suggestions to test consumers about how to select and use test products. The Standards seek to combine business with scientific norms, and are only sometimes successful. Consider, for example, the Standards' observation that "[l]arge sales make research financially possible." 1974 STANDARDS, supra note 219, at 7. Under ideal scientific circumstances, of course, research would precede "large sales," but the Standards neither require nor suggest this. See also supra note 326 and accompanying text.
332. Fincher, supra note 73, at 493.

The APA is certainly no worse than other professional organizations in taking principled stands that may adversely affect some of its members. In fact, it may be better than most in this regard and, from time to time, it has taken just such difficult and unpopular stands. However, the point is that it cannot consistently be relied upon to do so. Cf. A. Melore, Lawyers, Public Policy, and Interest Group Politics (1977) (discussing the interests advanced by the American Bar Foundation).
fectiveness and reported quality of effectiveness [evidence].”

334 Confronted by assessment consultants who often hold advanced behavioral science degrees, employers are in no position to adequately evaluate the tests being sold to them. Furthermore, some test psychologists have observed, “employers are primarily interested in developing a selection device which will withstand court scrutiny, rather than one which is ‘truly’ valid.”

335 Employment testing is not a science. Indeed, it has become very much a business. Especially in determining cut-off scores, deciding the tolerance for certain kinds of errors in prediction, and judging the practical and legal acceptability of employment tests, professional testers have no decisive claim to expertise:

If one judge prefers to emphasize one outcome and a second prefers to emphasize another, there is no scientific or logical basis to defend one emphasis over the other. . . . This reiterates one of our fundamental theses: when mathematical formulas are used as guides to policy making they carry hidden value judgments that the decision maker might be unwilling to accept if he considered them explicitly.

336 Employment test regulation must force explicit utility statements on test users and scrutinize testing practices from the perspective of social and constitutional values. Professional values are not sufficient.

C. A Federal Testing Agency: Direct Regulation and Centralized Validation Data

Although the current APA Standards do not contemplate violations or impose sanctions on professionals for bad testing practices, psychologists have long been aware of test abuses. Soon after testing

334. Blumenfeld, Thurman, & Peterson, supra note 273, at 50. See also supra notes 271-73 and references cited therein.


336. L. CRONBACH & G. GLESER, supra note 230, at 128. Given the credentialing norms of these professions, psychological and industrial decisionmakers are likely to have done well on these and similar tests themselves. Thus, it may be easier for them to overvalue the tests and have confidence in their validity. For a cogent plea to evaluate the social impact of testing, and to do so with standards whose force and logic derives from outside the tests and measurement establishment, see Scriven, the Logic of Judgment in Evaluation and the Law: Making Hard Decisions with Soft Data, in EDUCATIONAL TESTING SERVICE, EDUCATIONAL MEASUREMENT AND THE LAW (1978).

337. See, e.g., supra note 39. It should also be noted that the APA was concerned with the standards of psychological testing as early as 1896. Some twenty years later, a resolution was passed to the effect that “this Association discourages the use of mental tests for practical psychological diagnosis by individuals psychologically unqualified for this work.” There was, however, no mechanism provided for the enforcement of the resolution, and in subsequent years several APA committees were appointed to investigate the feasibility of certification for persons who administered such tests. In 1923, one committee specifically recommended certification in Educational and Industrial Psychology, but the plan was rejected. In 1927, a modest program of certification in Consulting Psychology was discontinued by vote of the Association. At the same meeting, members failed to act on a motion that would have established an APA Committee to
became "big business" for psychology, concern was voiced over test developers and publishers who frequently failed to report the limitations of the tests that they marketed.\textsuperscript{338} From time to time, professionals also have suggested that more stringent internal regulation of test practices be imposed. In the aftermath of Griggs, one of them suggested that the APA consider "professional sanctions against publishers and psychologists who use invalid tests to perpetuate racial discrimination in employment."\textsuperscript{339} The suggestion was not followed. Another psychologist proposed that the APA and major test companies exercise a more general regulatory function in the use of employment tests by promulgating "minimal prerequisites that a company or institution would have to complete before additional test materials would be sent."\textsuperscript{340} Unfortunately, neither the APA nor the test companies possess the institutional competency to perform such a function. They have neither the monitoring capacity nor the power to compel compliance with test regulations or standards. Moreover, their interests, and those of their clients and constituents, are often at odds with the demands of effective test regulation.

What is needed instead is a national center for test validation that would carefully scrutinize and certify tests, supervise the design of predictive validation studies, grant approval for cross-validation (use of tests validated on slightly different jobs or populations), make appropriately validated situational and work-sample tests available to employers and personnel consultants, develop and impose the appropriate utility functions to govern employment testing decisions, and serve as a clearinghouse for test information by accumulating, compiling, and evaluating data on employment tests. It might also oversee the licensing of test administrators whose current status is virtually

\textsuperscript{338} Holtzman, \textit{Some Problems of Defining Ethical Behavior}, 15 AM. PSYCHOLOGIST 247 (1960). Consider also the problem noted by Bollenbacher, \textit{Review of Standards for Educational and Psychological Tests}, 12 J. EDUC. MEASUREMENT 55, 56 (1975): "In general, however, it is the opinion of this reviewer that the Standards in their present form will probably have a minimal influence on many school people and others who use tests routinely in much of their work. Why? Because many teachers, administrators, and even counselors will continue to be unaware that any such body of standards exists." Finally, a National Academy of Sciences Committee noted that "most test users are not members of the organizations that have taken the lead in setting professional standards and are thus not subject even to their mild form of quality control." \textit{Committee on Ability Testing, supra} note 254, at 20.


Unlike the EEOC, which must investigate individual grievances on a case-by-case basis, a federal testing agency could take a more "proactive" role in the stringent regulation of tests, enforcing standards on a much broader and more consistent basis. With the assistance of the federal courts, this agency could also restructure the employment test practices of violating companies to insure full guideline compliance. Abraham Chayes has noted the inadequacies of traditional civil lawsuits for competently addressing the complex issues and implementing the special orders that issue from complicated polycentric public interest cases. Many employment testing cases have precisely the features he describes, and a federal testing agency could assume the monitoring and implementation responsibilities after a decision had been rendered in a testing case.

"Proactive" regulation also avoids the pitfalls of reactive law, in which legal response is conditioned upon demonstrations of abuse and injury. Particularly in light of the "psychologic" of testing, which encourages the internalization of test failure, there is no reason to believe that all aggrieved parties will pursue available, traditional litigation remedies. A federal testing agency could help curtail uncontrolled test distribution and foreclose much test misuse before it damaged any test-takers.

In a sense, the formation of a testing agency would shift the tester's burden of proof to an earlier stage in the validation process. Tests could not be used at all until they had been properly validated. The converse logic—that tests are used until adverse impact is demonstrated, or until failure to validate them precipitates litigation has dominated test practices until now. By controlling test certification and distribution, however, a federal testing agency could prevent rather than simply remedy test abuse. Such an agency also would likely reduce the amount of testing litigation that reaches the courts. By collecting and centralizing test data, and certifying tests for distribution, the

341. A few years after the Civil Rights Act of 1964 was passed, one test psychologist wrote that "[the beginning of ethics is technical competence, and calls are being heard for a psychological bureau of standards to help define competence in this sensitive area]" Guion, supra note 134, at 209. It is time those calls were heeded, albeit on a far wider and more independent scale than the "internal housekeeping" then contemplated.


345. See 44 Fed. Reg. 12,001 (1979), concerning the advance review or test procedures by existing enforcement agencies: "Enforcement agencies will not review validity reports except in the context of investigations or reviews."
agency would make standards more clearly defined and the outcome of various testing practices more predictable. Invalid tests could be prohibited from use and hence would not generate litigation.

Test psychologists often complain that the rigors of predictive validation cannot be met with small subject samples. A test agency could remedy this problem by pooling data under systematic and controlled conditions. When utility functions are employed in which false negatives must be calculated and weighed heavily in decision theory models of fair employment, convenient concurrent validation strategies cannot be pursued.346 Elaborate evaluative models increase the need for centrally coordinated collaborative studies so that the costs of such studies can be more fairly distributed and the results more systematically combined. Arguments about the inconvenience and infeasibility of predictive validation would thus lose their force. Examination of the racially discriminatory effects of testing would become feasible with these larger samples as well.

Agency-supervised validation studies would eliminate the questionable, makeshift validation procedures that have sometimes surfaced in testing litigation. The agency could impose some degree of quality control on the studies as they are conducted, including the selection, training, and supervision of persons who perform validation studies.347 The relative absence of clear standards and the presence of varying degrees of sophistication in the judicial review of employment test issues could be remedied by developing a coherent and uniform national policy, created and administered by an agency specializing in the critical evaluation of test instruments.348


347. See 44 Fed. Reg. 12,002 (1979), to the effect that a validity study "may be performed by any person competent to apply the principles of validity research." There are currently no legal or professional definitions of such persons.

348. Two very real dangers associated with the creation of such an agency must be acknowledged. The first is that the specialized nature of agency work makes it likely that its staff will be dominated by former and future employees of the test industry, to a degree that will compromise regulatory action. While this problem plagues all governmental regulatory bodies, it does not seem to have impeded the work of the most closely analogous agency—the EEOC. Secondly, it is possible that the agency certification or licensure of certain tests could invest them with far more legitimacy and importance than they deserve or might otherwise obtain. This danger must be carefully evaluated against the costs of unregulated, or reactively regulated, test use.
D. Development of New Assessment Models

As employment tests continue to be rigorously evaluated in terms of predictive validity and practical utility, new models of assessment must be developed for employers who still seek personnel screening devices. In light of existing evidence, criterion or work sampling tests should always be preferred to more convenient, but less meaningful, alternatives. Through job analysis, test developers must determine what tasks specific jobs actually require, and structure these major work components into the test itself. Properly done work samples help to insure the job relatedness of the tests. Sampling must be done on a representative basis from the complete array of required job tasks, and the tendency to sample those behaviors most amenable to convenient forms of test assessment—like verbal ability—should be resisted. Racially-biased or culturally-bound work samples—ones that rank the importance of work tasks according to the majoritarian preferences of test developers rather than to the tasks' actual significance to the job in question—are scarcely improvements over standard employment tests and should be rejected as well.

New methods of assessment that are highly sensitive to changes and improvements in performance should be used. Unfortunately, "[o]ne of the hidden prejudices of psychology, borrowed from the notion of fixed inherited aptitudes, is that any trait, like racial prejudice, is unmodifiable." Skills, however, are modifiable, and new testing models should be far more sensitive to the adaptability of human behavior by incorporating practice or trial sequences into the performance measures. Indeed, with new assessment models we might spend less time trying to uncover "true" aptitude or ability and more time training applicants to criterion level performance. In general, these new models might be far less static in nature than traditional ones. That is, rather than assessing ability or aptitude at a single point in time, and basing employment decisions on the outcome, they might take a more sequential approach in which assessment sessions are regarded essentially as investigatory attempts at gaining more information before terminal decisions are made. Periods of trial or


350. For favorable reviews of job assessment centers in which elaborate simulations of actual working conditions are employed, see Howard, An Assessment of Assessment Centers, 17 Acad. MGMT. J. 115 (1974):

A unique contribution of assessment centers is the inclusion of situational tests in the assessment battery. The rationale behind using such exercises is that they simulate the type of work to which the candidate will be exposed and allow his performance to be observed under somewhat realistic conditions. Contrary to the aptitude test approach, samples, not signs of behavior are used for prediction.

Id. at 117.

351. McClelland, supra note 150, at 8.
probationary work, supplemented with on-the-job training, might be used instead of pre-employment screening. It is also possible to take an adaptive rather than fixed treatment approach to jobs and working conditions themselves. That is, rather than attempting to select people who “fit” the job, employers might restructure jobs more to fit employees.

Even alternative and innovative assessment strategies, however, must be pursued along with ongoing predictive validation studies. Moreover, validation studies must be conducted with groups that are representative of the larger applicant pool, including minorities. Holtzman notes that often “[i]n industries where blacks have been discriminated against until recently, there is an insufficient amount of evidence from previous research to defend the test procedure against charges of bias.” But it is precisely in these industries, where the atmosphere and effects of past discrimination may persist, that disparate impact employment tests must be scrutinized most carefully. The costs of conducting these validation studies, as well as the costs of developing new assessment models, must be borne in part by employers. Despite numerous protests to the contrary, there may be both individual and distributive justice in proposals that require employers to shoulder some of the burden in establishing fair employment policies.

Employers

---

352. This is to be distinguished from the procedure employed by defendants in Washington v. Davis. These applicants were screened first on the basis of test scores that attempted to predict training school performance. Every person whose predicted performance was above a certain level was accepted into police training and given additional training until meeting the criteria of successful training performance. No data were or could have been collected on the ability of rejected applicants to meet the criteria of training performance with or without the additional training.

The armed forces have experimented successfully with remedial training programs given to servicemen who perform poorly on classification tests. See E. Ginzberg & D. Bray, The Uneducated (1953); Goldberg, Psychological Procedures Employed in the Army's Special Training Units, 1 J. CLINICAL PSYCHOLOGY 118 (1945). See also Kirkpatrick, supra note 132, at 32-33; Goldstein, Training in Work Organizations, 31 ANN. REV. PSYCHOLOGY 229 (1980); Sharf, Do's and Don'ts in the Employment Process: The Current Status of Testing, 1 Employee REL. L.J. 579, 588 (1976); Siegel & Bergman, A Job Learning Approach to Performance Prediction, 28 PERSONNEL PSYCHOLOGY 325 (1975).

353. See L. Cronbach & G. Gleser, supra note 230, at 143:

American testers have placed themselves at the disposal of institutions wishing to assign men to predetermined treatments, and have pointed with pride to the fact that they often can raise the average output by careful selection. But nearly equal gains can often be attained by another branch of personnel technology which refuses to accept the treatment as given. The job simplification expert and the human engineer seek to fit the job to unselected men. The greater their success the less the value of selection. The tester has failed to realize that he is competing with the treatment simplifier. And the latter's method is often the more economical, for his changes may be made permanent while the tester must evaluate new employees forever.

See also Cronbach, supra note 129, at 48.


355. See United Steelworkers v. Weber, 443 U.S. 193 (1979). R. Wasserstrom, supra note 3, at 79 n.10, has suggested that the doctrine of unjust enrichment could serve as the rationale for
are the direct beneficiaries of an economic and political system premised on promises of fair and equitable treatment. The continued viability of this system may depend on its ability to match current reality to the rhetoric of fairness. Moreover, even though employers may not have consciously discriminated themselves, they are likely to have been the indirect beneficiaries of discrimination, in terms of the superior educational and occupational opportunities afforded them. It is not inappropriate, then, to require them to participate in the restoration of racial equity in the workplace.

E. A Reexamination of Testing, Meritocracy, and Institutional Racism

Public and professional debates over testing often implicate a "meritocratic" value system in which the most "qualified" are thought to deserve or be entitled to the greatest opportunities and rewards. Employment tests are often defended, and legal restrictions on the use of test scores in employment decisions are challenged, in the name of this meritocracy. Two notions are at the basis of the meritocratic claims: (1) an economic efficiency or "business necessity" argument that the most qualified should be afforded the greatest opportunity and reward because they will be the most productive workers, and (2) a quasi-moral argument that the most qualified "deserve" or have a superior moral claim to opportunity and reward merely by virtue of their greater talents and abilities. In most cases, these two arguments are conflated and asserted uncritically in what becomes a philosophically and psychologically confused concept of "merit." A fair employment policy that seeks to gain widespread public support, however, must disentangle the various aspects of this concept. Several issues should be kept in mind during this process of disentanglement.

compensatory remedies in cases in which the person from whom a benefit is taken has not necessarily caused harm to another. His suggestion is worth considering in this related context.


357. See, e.g., Goldman, supra note 3. After a painstaking defense of the ultimate justice in "hiring by competence," he offers the following "fact of life":

Some people simply do not have the native intelligence, and any system of distributing jobs in which efficiency is an important consideration will forever bar these individuals from positions in the upper echelon. I maintain that if this is sad or seems unfair, it is nevertheless a fact of life which cannot be justifiably improved by alternative rules of hiring.

Id. at 28.

358. For example, this notion finds expression near the end of Justice Powell's opinion in Bakke. He writes that "[f]airness in individual competition for opportunities, especially those provided by the State, is a widely cherished American ethic." 438 U.S. 265, 320 n.53. Thus, the unfairness to non-minority applicants of the special admissions program challenged in Bakke is that "[n]o matter how strong their qualifications . . . they are never afforded the chance to compete with applicants from the preferred groups for the special admissions seats." Id. at 320.
The first is that attempts to equate "merit" with ability often overlook the imprecision and narrowness with which ability is typically measured. Employment testing is a highly imperfect and error-ridden technology. Ironically, those who defend tests on meritocratic grounds often vehemently resist the imposition of legal standards that require careful scrutiny of how well tests measure what they purport to measure. Before the connection between "merit" and business necessity can be made plausible, however, there must be a convincing demonstration of the relationship between job performance and the "merit" in question. Test scores that reflect nothing more than narrow competencies in test-taking do not provide such a demonstration. In this context, of course, business necessity should not be confused with mere business convenience. Tests are easy and convenient to administer and they do produce a distinct ordering of persons. But before this ordering rises to the level of a meritocratic concern, it must be meaningful and valid. Too often, it is not. 359

Similarly, when the concept of "merit" is used to convey a notion of human worth or moral deserving, it should not be confused with the ability merely to answer standardized test questions or to perform on paper-and-pencil measures of aptitude and skill. Were the rejected black applicants to the District of Columbia police force in Davis truly devoid of "merit" because they were unable to demonstrate knowledge of words like "impertinence" and "promontory"? 360 Somehow the test industry has interposed its technology into subtle and complex decisions about human value and worth. But there is no obvious reason why persons who score well on employment tests necessarily "deserve"—in moral terms—anything more than the high test scores that they receive. Especially when merit is defined in such a way that dominant culture characteristics and preferences are incorporated to the exclusion of minority ones, the concept cannot and should not be asserted as a moral absolute. Explicit recognition of the special merit that racial minorities bring to the workplace by virtue of their unique perspectives and experiences is absent from test-based selection.

The common equation of test performance with "merit" is usually premised, however vaguely, on a dispositional model of human behavior. That is, the test results are usually taken to reflect some more or

359. See supra notes 172-227 and accompanying text. See also Brickman, supra note 132, which provides some empirical support for the proposition that the perceived fairness of a selection procedure affects our willingness to accept alternative procedures like lotteries or random selection.

360. The test used to screen applicants in Davis is reprinted in 512 F.2d at 967-76. Test advocates sometimes bristle at the criticism of individual test items, and argue instead that a test's sophistication is reflected only in the sum total of all its items. But this argument attaches properties to the tests that they simply do not have. The whole of these tests really amounts to nothing more than the sum of their individual item-parts.
less important and stable characteristic of the person in question. These characteristics are not always regarded as intrinsic or innate. Sometimes they are taken to represent a lifetime of preparation, effort, and hard work. Persons are then thought to deserve their rewards because they have "earned" them. Yet test scores are so highly correlated with variables over which individuals have literally no control—like race and socioeconomic class—that it is quite unclear how many people truly can be said to have "earned" their scores. When test scores are seen more as the product of narrow test-related competencies that are not necessarily useful in predicting important behaviors like job performance, they will less often be taken as measures of deep-seated abilities or traits. And when both test and job skills are seen as more situationally-specific and highly modifiable, they will be less often confused with essential "merit."

The current controversy over the effect of "coaching" for certain kinds of tests provides a case in point. If the intellectual "aptitudes" supposedly measured by tests are relatively fixed and enduring qualities of the individual, then test scores should be little affected by training sessions designed to familiarize persons with question format and material. Short-term attempts to elevate test scores, and thereby enhance scholastic "merit" developed over a life-time of educational experience and effort, should fail. This has been the position of the major educational test publisher, Educational Testing Service. On the other hand, if these aptitudes, like most human behavioral characteristics, are highly modifiable and subject to change by immediate or recent experience, the courses might be quite effective. The available evidence is decidedly in favor of the latter view. Some coaching courses have been highly effective in producing dramatic increases in basic "aptitude" scores. In what sense, then, are persons who have had the opportunity to attend such courses more "meritorious" than those who could not attend?

Even if ability could be measured in an accurate and meaningful fashion through testing, another problem is raised by the meritocratic claims with which tests are defended. These claims are often asserted in such a simplistic and absolute fashion that other legitimate competing concerns and values are ignored. For example, one psychologist writes that "if the term 'fairness' has any objective meaning at all, it must mean selecting the best people... thus 'fairness' is nothing more
or less than simple validity.” Of course, social and legal concepts of “fairness” are far more complex than that, and legitimately include important considerations of equal opportunity and racial justice. Indeed, we tolerate gross deviations from the meritocratic ideal in other sectors of our society, often at the very highest levels, presumably in the name of other legitimate values against which meritocratic claims must be balanced. Employment testing seems a very curious place at which to rigidly insist on a standard that is rarely enforced elsewhere, particularly in light of the imprecision and unreliability with which the standard is assessed in the employment test context.

Sometimes, however, tests are defended in explicitly democratic terms, as the very vehicles by which equality of opportunity may be obtained. The values of the meritocracy, it is suggested, are highly compatible with those of racial justice. Tests, in this context, are offered as impartial and objective instruments for overcoming the disadvantages of status, caste, and race: “The tests couldn’t see whether the youngster was in rags or in tweeds, and they couldn’t hear the accents of the slum. The tests revealed intellectual gifts at every level of the population.” Indeed, survey data suggest that most Americans believe tests to be “color blind,” constituting universal standards of excellence that give minorities an equal chance that the overt racism of an earlier period denied. “Minority groups,” we are told, “should be favorably inclined to the use of ability tests, since tests constitute a universal standard of competence and potential.”

Against this public image of tests, and the democratic virtues that their meritocratic use supposedly promotes, is the social reality of testing to which I earlier adverted. The negative effects of testing are not merely arbitrary and capricious; often they are systematically discriminatory. Many tests operate with uncanny regularity to exclude minorities from a whole range of educational and economic opportunities in this society. The notion of “merit” advanced by test scores correlates extremely well with race and with class. For example, on National Merit Scholarship tests, in a program that prides itself on its exclusive

364. Darlington, A Defense of “Rational” Personnel Selection, and Two New Methods, 13 J. EDUC. MEASUREMENT 43, 44 (1976). (Emphasis in original.) The lengthy debate over statistical fairness conveys the misleading impression that fairness is simply a matter of psychometric theory. Typically the debate also assumes much of what is at issue in employment testing cases. That is, it involves statistical arguments over how to make the best of what are often very bad selection devices.


367. Brim, supra note 318, at 127.

368. See supra notes 130-41 and accompanying text.
concern for "intellectual excellence," test scores correlate almost perfectly with estimates of parental income. Or consider the U.S. Civil Service Commission, often touted as an example of purely a competitive merit system. The Commission was recently sued for using an employment test for upper-echelon positions that was passed by white applicants at a rate that was nearly nine times the pass rate for blacks. Once these supposedly "pure merit" systems are carefully scrutinized, it becomes clear that many of them operate with inexorable logic to preserve existing distributions of privilege and advantage. Far from neutralizing the effects of pre-existing inequalities, many testing programs amplify these effects through the seemingly neutral principle of "merit."

In this sense, the continued emphasis on test scores as the primary indicia of merit constitutes a powerful form of institutional racism—the use of seemingly neutral principles that consistently operate to the detriment of minority interests. The confusion of test scores with human worth or merit perpetuates racial inequality. Testing becomes the vehicle by which differential opportunity and reward can be explained and justified because it functions to translate discriminatory treatment into ostensible matters of comparative worth and deserv- eness. This process is facilitated by the still generally shared public belief in the objectivity of these devices, and the lack of general awareness of the nature of the massive discriminatory impact of testing. Naive

371. See, e.g., Kavruck, Thirty-Three Years of Test Research: A Short History of Test Development in the U.S. Civil Service Commission, 11 AM. PSYCHOLOGIST 329 (1956); Macy, supra note 156.

It is of some historical interest that social scientists were among the very first to press for civil service "merit" systems. The American Social Science Association—precursor of various contemporary professional social science organizations—adopted merit-based civil service as its cause celebre in the nineteenth century. Historians have interpreted this as a somewhat self-serving attempt by early social scientists to redefine "authority" in terms of "competence." Systems that formalized and institutionalized the certification of merit not only seemed reasonable and badly needed but were also quite useful to this credentialled and "competent" group. See T. Haskell, The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth Century Crisis of Authority 115-21 (1977).

373. A fine working definition of institutional racism is offered by a former president of the American Psychological Association: "This term labels characteristics and processes of the legitimate ongoing institutions of society that tend to perpetuate the second-class status of Negroes, and of other disadvantaged minorities. To be sure, institutional racism finds support in personal bigotry, perhaps especially in its more subtle and unacknowledged forms. But, more importantly, it finds its major support in all of the many honorable values and reasonable justifications for continuing existing institutions such as they have been." Smith, Racism, Education, and Student Protest, ILL. SCH. J. 207, 211 (1969).
faith in the test instruments themselves is supported by the larger technological context in which they flourish, and the belief among many persons that any numbers—however unreliable—are better than none.

In sometimes arbitrary and discriminatory fashion, tests do simplify employment decisionmaking. Tests enable ranking and numbering systems to be developed, that provide convenient mechanisms for allocating opportunities and rewards and they act to limit the sheer number of candidates who will be considered for jobs or promotions. Indeed, some critics have suggested that the latent function of employment tests and other selection criteria is to artificially restrict access to a very limited number of jobs, rather than to insure that only “qualified” workers are hired or promoted. But employment testing, by placing responsibility on the characteristics of individuals who supposedly “deserve” their differential treatment, draws attention from a system that generates patterns of inequality. When psychologists write that “400 years of slavery and other forms of economic deprivation have affected more than just test scores,” they are alluding to presumed cognitive and performance deficiencies among minorities. By focusing on alleged differences in individual performance, however, tests divert attention from the structural legacies of slavery and racism. A truly fair employment policy should be aimed at directly changing the structural impediments to equal opportunity. Fair employment policy awaits serious debate on several important questions—how best to eliminate structural barriers to equal opportunity, how to define and determine appropriate and meaningful job qualifications, and how to decide the role individual ability factors should play in a structurally fair employment system. A psychologically and philosophically confused concept of “merit,” one laden with surplus meaning, should not be interposed in this debate.

VI. Conclusion

Employment tests represent a technology whose logic is still too much assumed and too little examined, whose human costs and hurtful consequences are often ignored and discounted in the name of efficiency or cost-effectiveness, and whose widespread use may be more a marketing than a scientific accomplishment. When employers justify decisions that adversely affect the lives of thousands of people, and do so in the name of science, the burden of scientific proof is heavy indeed. In many instances, this is a burden that the test industry cannot meet. Employment testing in this country is far more a business than a sci-

374. See, e.g., G. SQUIRES, supra note 255: “[T]he function of increasing requirements is more to sort and select among candidates on the basis of their relative educational attainment and to restrict access to jobs rather than to insure that candidates are technically qualified” (at 186).

ence. Although founded on assumptions about human behavior that appear increasingly outmoded, employment tests continue to enjoy widespread success. There continues to be little connection between the quality of evidence that can be marshalled in support of employment tests and public confidence in their fairness and validity. Especially in the context of fair employment policy, decisions should not be permitted to rest on successful marketing techniques or the ideological predilections of those who make and use psychological tests.

The Supreme Court's decision in *Washington v. Davis* appears to represent a retreat from an earlier position that appropriately held employment testers to the highest standards of their profession. If a retreat is underway, then the Court has failed to see that continued use of disparate impact tests that are not truly job-related and racially fair can only be a pretext for discrimination or for a refusal to accommodate business convenience to fair employment policy. Neither goal should be acceptable.

Chief Justice Burger wrote in *Griggs* that "[d]iplomas and tests are useful servants, but Congress has mandated the common sense proposition that they are not to become masters of reality." Unfortunately, they tend to become precisely that. Even when used in conjunction with other techniques of evaluation and selection, the seeming objectivity of standardized test scores guarantees that they are given disproportionate weight in decisionmaking. Employment tests "[cloak] employment decisions with objectivity, [give] comfort to the employer who feels he now has the most qualified employee, and [protect] him from charges of nepotism or discrimination." Yet these may be their most treacherous features. If tests are not predictively valid, or are insufficiently job-related, then their objectivity is rendered a disguise, the comfort they give becomes false, and their protection becomes unjustified.

Imposing a more rigorous standard of regulation on the test industry may mean that standardized tests will be abandoned in some cases and their use seriously curtailed in others. But failures to demonstrate job-relatedness and test utility must weigh heavily in a national policy of fair employment. As the staggering physical and psychological costs

---

376. See supra notes 213-235 and accompanying text.
377. 401 U.S. at 433.
of unemployment are increasingly recognized, the legitimacy of any practice that systematically localizes these costs among certain groups of people must be scrutinized most carefully. It is time to reexamine the naive assumptions that individual merit and worth can be conveniently quantified and measured, and used as the exclusive mechanism by which opportunity can be allocated. In human terms, this assumption has become too costly.

It may still be debatable, in some legal contexts, for a society that has subjected minority groups to systematic discrimination in their living conditions and educational opportunities, to also countenance an employment policy that penalizes those same groups for the alleged absence of skills they were never provided or able to obtain. If the skills that these groups supposedly lack bear no demonstrable relationship to the jobs they are being denied, this injustice should be beyond debate.

---

380. See, e.g., Harvey Brenner's extensive study of the relationship between economic changes and "social costs" (as indicated by rates of mortality, suicide and homicide, mental hospital admissions, and imprisonment). Brenner concludes:

The most consistent pattern of relationship between national economic changes and each of the measures of social cost was demonstrated with unemployment rate. Unemployment plays a statistically significant role in relation to social trauma for each of the indices of social cost for virtually all ages, both sexes, and for whites and non-whites in the United States.