Volume 23
Number 3 *Spring 1989*

*pp.*213-265

Spring 1989

Minimizing Unnecessary Differences in Occupational Testing

Martin M. Shapiro

Michael H. Slutsky

Richard F. Watt

Valparaiso University Law Review

Volume 23 Winter 1989 Number 2

ARTICLES

MINIMIZING UNNECESSARY RACIAL DIFFERENCES IN OCCUPATIONAL TESTING

MARTIN M. SHAPIRO*

MICHAEL H. SLUTSKY**

RICHARD F. WATT**

I. INTRODUCTION

Since the early 1900s the United States has experienced an explosive growth in occupational licensing and certification.[1] Oregon enacted the first

* Professor, Department of Psychology, Emory University, and member of the Georgia bar. Dr. Shapiro served as a consultant to Golden Rule Insurance Company in the litigation discussed in this article and as Golden Rule's representative on the advisory committee established by the settlement.

** Partners in the Chicago law firm of Cotton, Watt, Jones & King, which represented Golden Rule Insurance Company in the litigation discussed in this article, as well as in other matters. The authors wish to thank Thomas D. Allison, Wesley Kennedy, Thomas Hancuch, Elaine Hale and Gillian Siegel for their assistance.

1. Although the terms "licensing" and "certification" are at times used interchangeably, there are distinctions. Licensing is "the process by which an agency of government grants permission to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency necessary to ensure that the public health, safety, and welfare will be reasonably well protected." B. Shimberg, *Occupational Licensing: A Public Perspective* 19-20, (ETS 1980) (quoting HEW, *Credentialing Health Manpower* 4 (HEW

barber licensing law in 1899, since then every state has adopted licensing requirements for dozens of occupations.[2] A 1980 government study found that more than 800 occupations and professions were regulated by state law.[3] In Illinois, for example, the Department of Registration and Education tests and licenses those who wish to engage in more than fifty occupations, including would-be teachers, nurses, real estate brokers, funeral directors, embalmers, land surveyors, shorthand reporters, and polygraph operators.[4] In addition, the Department of Insurance tests and licenses insurance agents and brokers.[5] With respect to teachers alone, nearly every state has adopted testing requirements as an element of educational "reform," assertedly to improve the quality of teaching.[6] Occupational testing is now so pervasive that there is little reason to expect it to diminish in the foreseeable future.

Private test developers have grown to meet and stimulate the demand

---

Publication No. (OS) 77-50057 1977)). Thus licensing is performed by a governmental agency and concerns minimum competency levels.

Certification, on the other hand, "is the process by which a governmental or non-governmental agency or association grants authority to use a specified title to an individual who has met predeterminated qualifications." *Id.* Governmental certification requirements, unlike licensing, do not determine whether a person may engage in a particular occupation, but rather whether he may use a title, such as certified public accountant. Certification standards are often above the level of minimum competency. *Id.* at 21-22. Non-governmental certification is frequently administered by professional or trade groups as a means of recognizing persons who have met certain specialized standards of practice. *Id.*

Because the principles discussed in this article apply generally to tests used for purposes of licensing or certification, the term "licensing" is used to include certification, except where the context indicates otherwise.

2. B. Shimberg, *supra* note 1, at 4 (citing K. Greene & R. Gay, *Occupational Regulation in the U.S.* (Employment & Training Admin., U.S. Dept. of Labor 1980)).

3. *Id.* The study did not include federal or local occupational regulation. *Id.*

4. R. Stackler, *Occupational and Professional Licensing Law* 57 (Ill. Inst. for Contin. Legal Educ. 1980).

5. Likewise, in New York, various agencies are responsible for testing and licensing persons seeking to engage in a variety of occupations. The Education Department administers the admission to and practice of some 28 professions, including nursing, landscape architecture, engineering, dentistry, social work, shorthand reporting and occupational therapy. N.Y. Educ. Law §§ 6500-6515 (Consol. 1985). The New York Public Health Department, under the direction of the Public Health Commissioner, tests and licenses funeral directors, undertakers and embalmers. N.Y. Pub. Health Law §§ 3420-3422 (Consol. 1985). The Superintendent of Insurance oversees the licensing of insurance agents and brokers. N.Y. Insur. Law §§ 1102-1109 (Consol. 1985).

6. J.T. Sandefur, *State Assessment Trends*, in 7 AACTE Briefs No. 6 at 12 (Aug. 1986); G.P. Smith, *Unresolved Issues & Developments in Teacher Competency Testing*, 8 Urban Educator No. 1 at 1, 2 (Fall 1986)(citing J.T. Sandefur, *State Assessment Trends*, 6 AACTE Briefs No. 2 at 21-23 (ERIC No. ED 260-115 Mar. 1985)). *See also* Flippo, *Teacher Certification Testing Across the United States & A Consideration of Some of the Issues* 4-5, presented at Annual Meeting of the American Educational Research Association (ERIC No. ED 260-115 Mar. 31-Apr. 4, 1985).

for occupational tests. Educational Testing Service (ETS), which styles itself "the nation's leading testing organization,"[7] is a major vendor of occupational tests. In 1973, ETS established a Center for Occupational and Professional Assessment (COPA)[8] to develop and administer licensing and certification examinations for government and quasi-public organizations. Occupational testing is big business: the ETS multistate real estate agents' licensing examination alone, used by twenty-five jurisdictions, has tested nearly two million applicants.[9] ETS also markets a multistate electrical licensing test, a certification examination for the Professional Golfers Association, a test for certified information systems auditors, and an energy auditors' certification program.[10] In 1987 ETS added a test to certify the competence of aerobics instructors.[11]

A principal allure of "standardized" tests is their purported objectivity and absence of bias. Given the proliferation of such tests for occupational licensing, the possibility of racial bias is a serious matter. How to define, identify, and remedy possible racial bias in testing confronts the psychological profession with a troublesome problem for which there are no agreed-upon answers. A closely-related issue faces lawyers and courts, namely, the standards applicable to claims of racial bias in occupational tests. The degree to which Title VII of the Civil Rights Act of 1964[12] bars discrimination in occupational licensing is unresolved. Nor is the standard clear for determining the requisite discriminatory intent[13] to establish a claim of racial discrimination in occupational testing under the Fourteenth Amendment.[14]

---

7. ETS, *The Center for Occupational and Professional Assessment - Licensing, Certification, and Assessment* 1 (1983).

8. *Id.* at 2.

9. ETS, *Real Estate Licensing Examinations, Bulletin of Information for Candidates* at 4 (1985).

10. ETS, *Report of the 1983 ETS Visiting Committee* at 9 (June, 1983).

11. ETS, 1987 *Annual Report* at 24.

12. 42 U.S.C. § 2000e.

13. In Washington v. Davis, 426 U.S. 229 (1976), the Supreme Court held that proof of discriminatory intent is required under the Fourteenth Amendment equal protection clause.

14. *See, e.g.*, Comment, *Challenges to Preemployment Tests After Washington v. Davis*, 5 HOFSTRA L. REV. 893 (1977). *See also* Bennett, *Reflections on the Role of Motivation Under the Equal Protection Clause*, 79 NW. U.L. REV. 1009 (1985); Binion, *"Intent" and Equal Protection: A Reconsideration*, 1983 SUP. CT. REV. 397; Comment, *Discriminatory Purpose and Mens Rea: The Tortured Argument of Invidious Intent*, 93 YALE L.J. 111 (1983); Clark, *Legislative Motivation and Fundamental Rights in Constitutional Law*, 15 SAN DIEGO L. REV. 953 (1978); Eisenberg, *Disproportionate Impact and Illicit Motive: Theories of Constitutional Adjudication*, 52 N.Y.U.L. REV. 36 (1977); Perry, *Disproportionate Impact Theory of Racial Discrimination*, 125 U. PA. L. REV. 540 (1977); Schwemm, *From Washington to Arlington Heights and Beyond: Discriminatory Purpose in Equal Protection Litigation*, 1977 U. ILL. L.F. 961; Ely, *Legislative and Administrative Motivation in Constitutional Law*, 79 YALE L.J. 1205 (1970).

This article considers both the psychometric and the legal problems arising from apparent racial bias in tests used for occupational licensing. The article focuses on a particular Illinois lawsuit—*Golden Rule Insurance Co. v. Washburn*[15]—which was settled after eight years in court. The Golden Rule Insurance Company, joined by several job applicants who "failed" the ETS-prepared examination, challenged the examination alleging that the test was racially biased and thus unfairly excluded minorities, particularly Blacks, from employment as insurance agents and brokers.[16]

The settlement agreement, executed in November 1984, imposed on the Illinois Department of Insurance and ETS a comprehensive set of procedures for assembling test questions into test forms. Those procedures generally require that, in each content area, questions with the least difference in Black and White passing rates have priority and that questions be pretested before being used for scoring purposes. In addition, the settlement provided for annual public reporting of detailed test statistics, including passing rates by different racial groups. The settlement is intended to minimize disparities between Black and White passing rates in insurance licensing tests in Illinois. The settlement thus represents a practical approach to a widespread problem, and, more importantly, there is significant evidence that the settlement works.

This article considers how the psychometric profession has dealt with the issue of racial bias in testing and why current approaches have not eliminated or significantly reduced racial differences in test performance. The article then examines the legal framework in which lawyers and courts address racial bias in occupational tests, concluding that current legal doctrine fails to protect adequately against test bias. Finally, in the absence of effective psychometric means for detecting and reducing test bias, and given the present unsatisfactory state of the law, the authors argue that the approach employed in the *Golden Rule* settlement represents a practical means for reducing racial differences in test performance.

II. Psychometric Techniques for Minimizing Racial Bias

Although every year millions of Americans take standardized tests which often significantly determine career paths and opportunities, test de-

15. Golden Rule Life Ins. Co. v. Washburn, No. 419-76 (Cir. Ct. Sangamon County, Ill., settled Nov. 20, 1984). *See* Golden Rule Life Insurance Co. v. Mathias, 86 Ill. App. 3d 320, 408 N.E.2d 310 (1980).

16. Over the past ten years, nearly a million candidates for insurance licenses have been required to take the ETS-prepared test. ETS *Annual Report* 24-25 (1983) (19 states and Bermuda); ETS, *1986 MILP Advisory Board Meeting Handout* 1-2 (Oct. 1986). In Illinois in 1985, nearly ten thousand persons took the life insurance portion of the examination. *See* ETS, *1986 Illinois Candidates for Licensure in Life Insurance and Accident & Health Insurance* (April 1986) Table I.

velopers remain essentially unregulated. It is ironic that an industry which claims it can devise objective measures of the educational and occupational abilities of so many individuals is itself subject to no meaningful measurement of how well and fairly it measures others.

The criterion for determining how well a test performs is test validity. While, on a theoretical level, there are various methods of validation, in practice the testing industry uses only one: content validity. As discussed below, that method and the existing procedures for detecting racial "bias" are not equal to the task.

The American Psychological Association (APA) has established *Standards for Educational and Psychological Testing.*[17] The *Standards* do not define "bias," although they do define two variants: "item bias" and "predictive bias." The *Standards* do not require, however, an absence of "predictive bias" as a necessary condition for test validity. In fact, the *Standards* mandate that predictive bias be investigated only when there is reason to suspect that it exists and then only if an investigation is technically feasible. Moreover, the APA offers its *Standards* simply as "guidance" for the exercise of "professional judgment." Similarly, the federal government's *Uniform Guidelines on Employee Selection Procedures*[18] do not define "bias"; they define "unfairness."[19]

Nonetheless, a common thread runs through the various definitions: "[A]n understanding of bias—as psychometric features that somehow misrepresent the abilities of one group—is expected to guide the detection of bias in particular instances."[20] In a general sense, bias is a failure to be straight; with respect to fabric, bias is a cut diagonal to a seam or pattern.[21] For present purposes, racial bias is the characteristic—admittedly difficult to identify and isolate—which causes a test to be other than straight with respect to a racial group. Members of that group do not perform as well on the test as do non-members, and the degree of disparity is not solely the result of differences in education or ability.

The real likelihood of possible racial bias in occupational testing is a matter of urgent concern. The problem, of course, is how to identify and eliminate the factors which make a test racially biased. Test developers and

---

17. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (1985) [hereinafter APA *Standards*].

18. 29 C.F.R. §§ 1607.1 - 1607.18 (1985) [hereinafter *Uniform Guidelines*].

19. Psychometricians debate the relationship between unfairness and bias and have proposed various definitions of each term. *See* R. BERK, HANDBOOK OF METHODS FOR DETECTING TEST BIAS (1982) [hereinafter BERK].

20. L. Shepard, *Definitions of Bias*, in Berk, *supra* note 19, at 25.

21. *Webster's New International Dictionary of the English Language* 262 (2d ed. unabridged 1952).

psychologists have not fully come to grips with racial bias. Traditional psychometric techniques do not work very well, and governmental and self-regulatory efforts have led to little more than a recognition that the problem exists.

A. Predictive Bias and Criterion Validation

Occupational testing is justified by the need to protect the public health, safety, and welfare by excluding persons who lack minimum qualifications from certain occupations. The unstated assumption is that there is a demonstrable connection between adequate performance on an occupational test and adequate performance on the job. If that connection cannot be shown, the test lacks validity: it does not test what it purports to test, and the public health, safety, and welfare are not served.

Psychometricians recognize what they call predictive bias: "the systematic under-or-over-prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance."[22] Predictive bias exists in a test (or, in the language of the federal *Uniform Guidelines*, the test is unfair) if members of one racial group characteristically score lower than members of another racial group and the differences are not reflected in job performance.[23] An example of the results of a racially unfair test is shown in Figure 1.

22. APA *Standards*, *supra* note 17, at 93. Under the *Standards*, "criterion performance" in the occupational licensing context would be a measure of job performance. *Id.*

23. *Uniform Guidelines*, *supra* note 18, § 16V, 29 C.F.R. § 1607.16V.

Figure 1

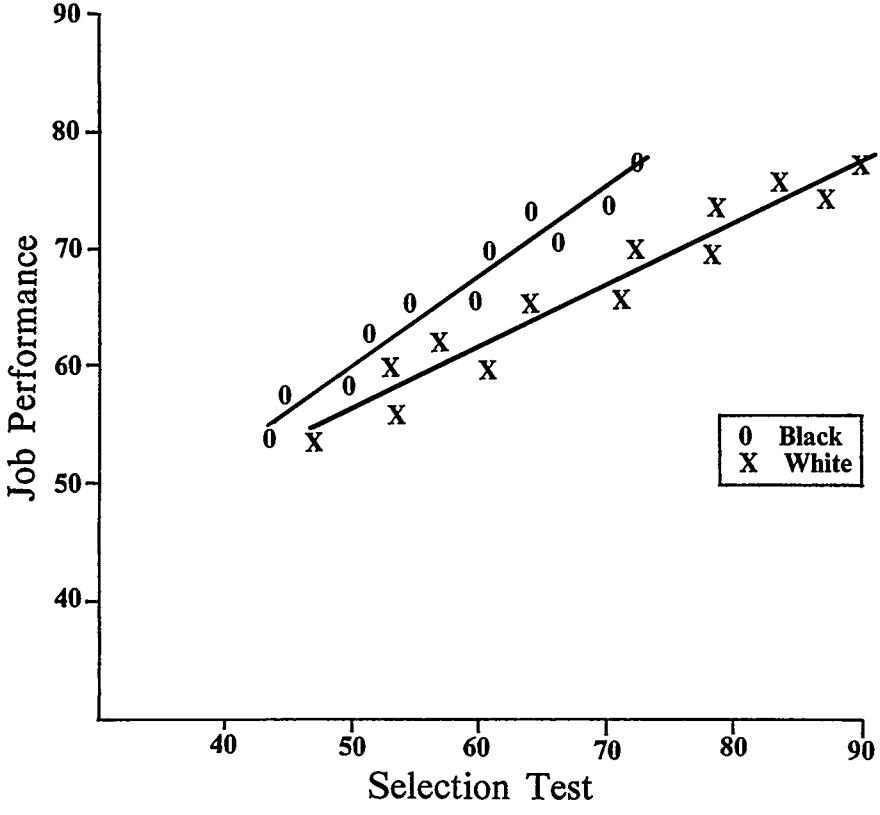
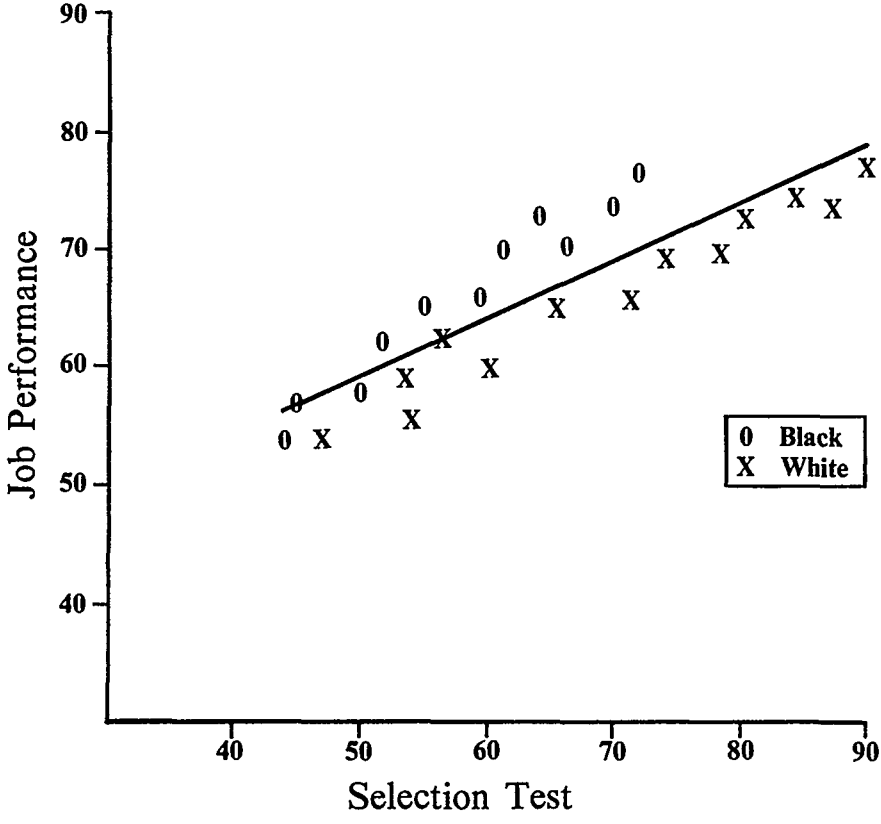


Figure 1 reveals that the average Black job performance is approximately equal to the average White job performance but, at any given level of job performance, Whites score higher on the test than Blacks. Predictive bias is present despite the fact that there is a reasonably positive relationship between test performance and job performance *considering all test takers*, both Black and White. The regression line which best fits the Black scores is not the same as the regression line which best fits the White scores.

When a regression line is drawn which best fits all of the points—both Black and White—and predicted job performance is based on that line, the bias becomes clear. Black job performance is under-predicted (actual Black job performance is higher than predictions based upon the regression line for all data points), and White job performance is over-predicted (actual White performance is lower than predictions based upon the regression line). Figure 2 graphically illustrates the bias. If the test were used to select

persons "qualified" to engage in the occupation (and to exclude those who failed), the result would be a disproportionate over-selection of Whites and a corresponding under-selection of Blacks.

Figure 2



In practice, matters are not so simple. Job performance can be measured only for individuals who have been hired. The measurement of job performance is necessarily restricted to only those individuals who scored satisfactorily on the test in question. It is impossible to determine the relationship between test performance and job performance for the entire set of individuals who originally took the test. Since the test itself determines who will be hired and therefore whose job performance will be measured, individuals who fail the test never have an opportunity to demonstrate their abilities on the job. Furthermore, subjective elements creep into virtually all evaluations of job performance. No matter how objective a test, its validity as a predictor depends in part on a non-objective evaluation of job performance.

Criterion validity—the correlation between test performance and job performance—is therefore difficult to achieve. A necessary, but not a sufficient, condition for criterion validity arguably should be the absence of predictive bias, but there is no firm consensus that this be so. There is general agreement that the developers and users of an occupational test have an obligation to come forward with evidence that the test is valid, but not that the evidence show that predictive bias was investigated and found absent.[24] Implicit in this failure to adopt a clear standard is the view that it would be impossible to demonstrate lack of predictive bias with regard to every conceivable group of people. The possible number of definable groups is simply too great. However, such a showing is not even required with respect to large minority groups, such as Blacks and Hispanics.

The APA *Standards* fail to require the absence of predictive bias as a necessary condition of criterion validity; the *Uniform Guidelines* do require consideration of less discriminatory alternatives[25] to a challenged test upon a showing that the test has "adverse impact" on a protected group but is nevertheless job-related. But "adverse impact" and "less discriminatory alternatives" have no defined meaning in the psychometric profession, even though the *Uniform Guidelines* provide a numerical measure of "adverse impact."[26]

---

24. *See, e.g.*, APA *Standards*, *supra* note 17, Standards 1.6, 1.7 and 3.10. The APA *Standards* provide:

Standard 1.6: When content-related evidence serves as a significant demonstration of validity for a particular test use, a clear definition of the universe represented, its relevance to the proposed test use, and the procedures followed in generating test content to represent that universe should be described. When the content sampling is intended to reflect criticality rather than representativeness, the rationale for the relative emphasis given to critical factors in the universe should also be described carefully.

Standard 1.7: When subject-matter experts have been asked to judge whether items are an appropriate sample of a universe or are correctly scored, or when criteria are composed of rater judgments, the relevant training, experience, and qualifications of the experts should be described. Any procedure used to obtain a consensus among judges about the appropriate specifications of the universe and the representativeness of the samples for the intended objectives should also be described.

Standard 3.10: When previous research indicates the need for studies of item or test performance differences for a particular kind of test for members of age, ethnic, cultural, and gender groups in the population of test takers, such studies should be conducted as soon as is feasible. Such research should be designed to detect and eliminate aspects of test design, content, or format that might bias test scores for particular groups. (*Conditional*).

Comment:

Although it may not have been possible prior to the first release of a test to study the question of differential performance and item bias for some groups, continued operational use of a test will often afford opportunities to check for group differences in test performance and to investigate whether or not these differences indicate test bias.

25. UNIFORM GUIDELINES, *supra* note 18, § 3B, 29 C.F.R. § 1607.3B.

26. *Id.* at § 4D, 29 C.F.R. § 1607.4D.

B. Content Validation

Criterion validation assumes the existence of a valid criterion (e.g., job performance) with which the predictor measure (e.g., selection test) can be correlated and thereby validated. Absent a valid criterion, there is no possibility of showing criterion validity. A valid criterion may not exist because the only available criteria are biased (e.g., subjective job performance measured by possibly biased supervisors) or because no acceptable criterion measure is available. The unavailability of valid criteria is common-place; with regard to licensing tests that unavailability is all but universal.[27] Nevertheless, each year thousands of prospective lawyers must take a bar examination, and thousands of aspiring insurance agents must take a licensing examination, notwithstanding the absence of any valid criteria for job performance. It is simply taken for granted that the tests predict who will and who will not perform adequately on the job.

Psychometricians in and out of the testing industry have devised another validation strategy, content validation, which is not dependent upon the existence of a valid criterion measure. Some psychometricians and psychologists, however, have expressed grave reservations about the entire concept of content validation.[28] Nevertheless, the paucity of criterion measures has created a perceived need for an alternative to criterion validation as a method of validation, and test producers have relied heavily on the content validation theory.

The APA *Standards* define content validity as the correspondence between the content domain of a test and the purpose of the test.[29] Content domain is defined as "[a] body of knowledge, skills, and abilities defined so that items of knowledge or particular tasks can be clearly identified as included or excluded from the domain."[30] The *Uniform Guidelines* are somewhat more restrictive in stating that content validity is inappropriate for

---

27. APA *Standards*, *supra* note 17, Part II, 10-11.

28. *Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*, 44 Fed. Reg. 11996, Question No. 54 (1979).

29. As set forth in the Glossary of the APA *Standards*, content-related evidence of validity is:

> evidence that shows the extent to which the content domain of a test is appropriate relative to its intended purpose. Such evidence is used to establish that the test includes a representative or critical sample of the relevant content domain and that it excludes content outside that domain. In employment selection testing, the content domain consists of tasks, knowledge, skills, and abilities associated with a job. In educational achievement testing, the content domain refers to the content of the curriculum, the actual instructions, or the objectives of the instruction.

APA *Standards*, *supra* note 17, Glossary at 90.

30. *Id.*

tests which purport to measure psychological traits or constructs.[31] Test developers seek to assure content validity by the procedures they use in constructing tests.

C. *Test Construction*

Ideally, to be consistent with the content validation model, constructing a test should involve: (1) conducting a job or task analysis; (2) identifying the job or task elements in terms of knowledge, skills, and abilities required by the job or task; (3) constructing test questions designed to measure the knowledge, skills, and abilities at the levels of proficiency required by the job or task; (4) reviewing the test questions for apparent lack of clarity, inaccuracies, and facial ethnic, racial or gender bias; (5) pretesting the examination on a sample of people to obtain data for analyses to detect signs of any such inadequacies or defects; and (6) assembling test forms for administration to future candidates. The process is inherently never-ending, if only because new test forms must be devised to reflect job changes and because the content of the test becomes known to test takers through breaches in security or systematic assembly of information by test preparatory course entrepreneurs or word-of-mouth reports by past candidates.

For purposes of content validation, a job analysis is conducted by actually observing work behaviors, interviewing job incumbents or supervisors, having job incumbents or supervisors fill out questionnaires, or convening discussion panels of job incumbents or their superiors.[32] The results are categorized and organized into job elements from which more macroscopic knowledge, skills and abilities can be identified, usually by a panel or panels of experts (job incumbents, former job holders, and supervisors) working with the job analyst or test developer.

The test developer typically then either convenes other experts who, after a brief instruction in question writing, draft questions, or requests experts in question writing to construct questions based upon the previously-determined knowledge, skills, and abilities associated with the job. In long-established testing programs, questions may be pulled from pools previously categorized by content.

---

31. The *Uniform Guidelines* provide:

> A selection procedure based upon inferences about mental processes cannot be supported solely or primarily on the basis of content validity. Thus, content strategy is not appropriate for demonstrating the validity of selection procedures which purport to measure traits or constructs, such as intelligence, aptitude, personality, commonsense, judgment, leadership, and spatial ability. Content validity is also not an appropriate strategy when the selection procedure involves knowledges, skills, or abilities which an employee will be expected to learn on the job.

UNIFORM GUIDELINES, *supra* note 18, § 14 C (1), 29 C.F.R. § 1607.14C.

32. S. BEMIS, A. BELENKY & D. SODER, JOB ANALYSIS (1983).

Following assembly, prospective test questions are reviewed by an expert panel for clarity and accuracy in order to ensure that for each question there is one and only one "correct" answer. Some test developers also convene a panel of diverse membership to analyze the questions for fairness and sensitivity. Such a panel permits the test developer to claim that its tests are reviewed for bias by members of different racial-ethnic groups of both genders. In practice, a "sensitivity review" relies on a most subjective measurement—each individual's opinion of what aspects of a test item might be offensive, unfair, or biased.

The assembled questions should then be administered to a try-out or pre-test sample. Although pre-testing is not always conducted even in large testing programs, there is general agreement that it is desirable.[33] A pre-test is the first opportunity to generate what are called "item analyses" or "item statistics." Additionally, the pre-test is the first opportunity to measure the test's possible adverse impact on specific gender or racial groups.

D. Item Analysis

After administering a test or pre-test, test developers subject the questions to an item analysis. An item analysis produces both statistics describing the response to each question separately and statistics regarding the relationship between responses to each individual question and responses to other questions. The other questions may be considered individually, aggregated as subscores, or aggregated as a total test score.

During the item analysis process, an item strip is generated for each question. The item strip records the number of persons and the percentage selecting each response alternative, as well as the number and percentage omitting an answer to the question (double answers generally being counted as omissions). Usually, the correct answer to the item is noted so that the percentage of test takers who answered the question correctly is readily ascertainable. Some item analyses also contain an additional index of item difficulty in the form of a standardized score calculated with reference to performance on other items of the test or performance on a set of items contained within another test form to which the current test is being equated. The typical item strip gives at least one correlation coefficient, measuring the relationship between performance on the particular item and performance on a larger set of items, usually performance on the total set of test items.

The correlation coefficient r is important because it measures a test's internal consistency. If a test is homogeneous in content, performance on

33. *See, e.g.*, ETS *Standards For Quality & Fairness* (1983), Tests & Measurement—Technical Quality of Tests, Guideline 7 at 12.

each item should be positively related to performance on each other item, and therefore performance on each item is related to performance on the test as a whole. If the test contains definable parts within which the items are homogeneous in content, performance on each item should be positively related to performance on the part of the test to which that item belongs. Generally, the greater the positive correlation with an appropriate aggregate score, the "better" the item; a negative correlation is an indication that an item, if not defective, is at least measuring something not consistent with the overall purpose of the test.

As a general rule, items whose correlation coefficients are below 0.3 are not considered "good" items.[34] Such items whose correlation coefficients are less than the benchmark or are negative may be flagged for review in order to determine if they are ambiguously worded or actually miskeyed with respect to the correct answer. Benchmarks are only general guideposts, and psychometricians are loath to adopt strict, inflexible standards for evaluating the quality of items. On the one hand, this failure to state precise decisional rules is a consequence of the fact that tests differ in their homogeneity and, therefore, in the expected values of r. On the other hand, the failure to state precise decisional rules reduces the analysis to a subjective exercise without an ascertainable standard for evaluating test quality, reliability, or validity.

Psychometricians view their item analyses as data for human judgment. But human judgment may serve as a mask behind which the failure to state a precise standard is hidden. Specifically, once an item is flagged, assuming a stated rule or rules for flagging, it is not clear whether there should be affirmative evidence for retaining the item or simply an absence of additional negative evidence for discarding it. Psychometricians do not state what presumptions are raised by, or what rules of persuasion apply to, items flagged by statistical criteria.

E. Item Analysis and Test Bias

There is widespread agreement that large positive item-test correlations are desirable, even if they are often excusably unattainable.[35] This consensus has potentially serious consequences. Consider the following example: Two groups of people, A and B, take a test. The individuals within each group differ from one another on the characteristic purportedly being measured. The test consists of 100 items, 80 of which accurately measure the characteristic under study, and 20 of which, in addition to measuring the characteristic within each group, are biased in favor of group A and

34. *See, e.g.*, R. Green, CTB/McGraw-Hill, *Comprehensive Test of Basic Skills: Technical Bulletin No. 2* (1977).

35. *See, e.g.*, A. ANASTASI, PSYCHOLOGICAL TESTING 211 (6th ed. 1988).

against group B. Assume that both groups have the same mean score on the 80 items. Obviously then, group A will outperform group B on the 100-item test which includes performance on the 20-item biased set. An item analysis for all test takers (ignoring A and B group membership) would show that the item-test correlation coefficients would tend to be highest for the 20 items in the biased set. This must be true because total scores are the sum of biased scores and unbiased scores, and only the biased items contain both the biased component and the characteristic being measured by all of the items. The correlation coefficients would be larger if group A was also superior to group B on the characteristic purportedly being measured.

The consequence is that the 20 biased items would be preferred for reuse, or (if the item analysis is part of a pre-test) those items would be chosen in preference to other items when the time came to construct new test forms. As successive generations of new forms were developed using item-test correlation coefficients based upon the entire population of test takers (ignoring group membership), the test would become progressively more biased in favor of group A over group B. If there were any basis for expecting group A to have a higher mean performance than group B, the entire difference between the groups could be attributed to that cause, even though much or all of the group difference observed would actually be attributable to biased items. Furthermore, new items which might be biased in favor of group B would not be used in current forms or reused in subsequent forms because they would have smaller item-test correlation coefficients. The test bias would perpetuate and enhance itself.

F. The Limits to Traditional Methods of Detecting Bias

Traditional psychometric methods designed to detect and measure item bias are all based upon a common methodology of conducting item analyses for separate groups of examinees rather than conducting one item analysis for the entire population of test takers. More precisely, the available psychometric measures of item bias do not measure item bias *per se* but only item bias relative to overall test bias. These methodologies can only detect whether a particular item is significantly more biased or significantly less biased than the aggregate of all the test items as a whole.

In practice, one cannot ascertain whether a test as a whole is biased without a criterion measure of known unbiased validity to which the test may be compared as a whole. Needless to say, if such a criterion existed, one would not rely upon content validation because criterion validation already would have been attempted. In reality there is no way, using traditional methods, to demonstrate that a content-validated test as a whole is unbiased. Conversely, there is no way to demonstrate that items other than the significantly deviant items are either biased or unbiased. Purging a content-validated test of its significantly deviant items only serves to make the

test more homogeneous with regard to bias, but not demonstratively biased or unbiased. The test as a whole is the standard against which individual items are judged, and discarding individual items which are flagged as deviant only makes a relatively minor adjustment to the self-justification of that standard. A heterogeneously distributed set of test items would be reduced in the range of its bias, but the mean bias would be little altered whether the mean is zero or large. A homogeneously biased set of test items would be unaltered in any way.

Even assuming that these traditional methods identify and flag for review items which may be biased, the review of such items is necessarily subjective: it consists of evaluating linguistic content and considering possible cultural explanations for the item behavior. There is no consensus regarding the standards for such a review. Alternatively, a scatter diagram could be constructed in which each test item is represented by a point numerically specified on one axis by the percentage of White candidates answering correctly and on the other axis by the percentage of Black candidates answering correctly. Deviant items could be identified as points which lie further than a specified number of standard deviations from the regression line, and all such items could be discarded.

Such a strict procedure is not employed and, even if it were, the selection of a replacement item would still remain subject to judgment. Human judgment would be the final determinant of the item content of the examination. No matter what procedure is employed to deal with flagged items, the remainder of the items—the items which lie proximate to the regression line—are implicitly accepted as having passed the bias test.

The existing statistical methods for detecting item bias are relative methods. They are restricted to comparing the bias of a particular test item to the bias of a part of the test or the whole test.[36] Such is the unsatisfactory state of the psychometrician's art, affording small comfort to those

---

36. As one authority has observed:

> *All statistical procedures devised to date are "relative" methods.* The present conceptualizations of bias in an item involve its "not fitting in" with some other group of items with which it had been thought to belong. This is true whether one uses matched or random samples and whether one uses internal or external criteria of ability. Thus bias is not necessarily some inherently "bad" characteristic of an item; it is dependent on the pool of items with which the particular item is being compared. In this respect, then, bias in the sense of unfairness to some group cannot be eliminated during test construction by rejecting items that do not meet a particular standard of comparability with other items. What the process can do is to improve the homogeneity of the test being constructed so that each item is yielding information parallel to that yielded by each other item. The final test must then be evaluated for fairness or bias in terms of the use to which test scores are put in decision making.

L. Burrill, *Comparative Studies of Item Bias Methods*, in BERK, *supra* note 19, at 173-74 (emphasis in original).

concerned with racial bias in occupational testing.

III. Statutory and Constitutional Standards

Two sources of federal law are available to those injured by racial bias in testing: Title VII of the Civil Rights Act of 1964,[37] and the Fourteenth Amendment.[38] Neither is particularly satisfactory, since courts, like psychometricians, are still groping. There are unresolved questions regarding the extent to which Title VII applies to licensing or to test developers. Likewise, the constitutional issues are unsettled: it is unclear to what degree the equal protection and due process clauses provide effective bases for attacking racial bias in testing. The result is that individuals adversely affected by test bias face formidable obstacles in court.

A. Title VII

Title VII prohibits employers from failing or refusing to hire any individual or discriminating "against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color. . . . or national origin. . . ."[39] Title VII contains the following caveat, however:

> . . . nor shall it be an unlawful employment practice for an employer to give and act upon the results of any professionally developed ability tests provided that such test, its administration or action upon the results is not designed, intended or used to discriminate because of race, color, religion, sex or national origin.[40]

The Equal Employment Opportunity Commission (EEOC), which, along with other federal agencies, enforces Title VII, issued interpretative guidelines,[41] known as the *Uniform Guidelines on Employee Selection Pro-*

37. 42 U.S.C. § 2000e.

38. Another basis for challenging licensing requirements, section 1 of the Sherman Act, 15 U.S.C. § 1, was effectively precluded by the Supreme Court in Hoover v. Ronwin, 466 U.S. 558 (1984). There the Court held that a suit against bar examiners for allegedly conspiring to limit the number of attorneys admitted to practice was barred by the doctrine of state action immunity from antitrust liability.

39. 42 U.S.C. § 2000e-2(a)(1).

40. 42 U.S.C. § 2000e-2(h). It is not a sufficient defense that a test was "professionally designed or developed" by a "professional" test developer. *See* Griggs v. Duke Power Co., 401 U.S. 424, 435-36 (1971).

41. The other agencies are the Department of Justice, 28 C.F.R. § 50.14 (1988); the Office of Personnel Management, 5 C.F.R. § 300.103(C1988); the Office of Federal Contract Compliance of the Department of Labor, 41 C.F.R. § 60-3 (1988); and the Office of Revenue Sharing of the Department of Treasury, 31 C.F.R. § 12.101 (1985).

cedures.[42] Under the *Uniform Guidelines* and Supreme Court decisions,[43] if a test or other selection procedure has an "adverse" impact on the basis of race, sex, or ethnic group[44] the employer must either eliminate the impact or show that the selection procedure is "job related."[45] Job-relatedness is established if the test has been validated for the purpose for which it is used.[46] Upon a showing of adverse impact, the burden shifts to the defendant to establish job-relatedness.[47] Under Title VII, a plaintiff proving ad-

---

42. *See supra* note 18.

43. *Griggs*, 401 U.S. at 431 (1971); Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975).

44. Section 4D of the *Uniform Guidelines* provides that "[a] selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5)(or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." 29 C.F.R. § 1607.4D. Section 4D also includes caveats to this "four-fifths" rule of thumb with respect to greater and lesser differences and with respect to statistical significance. The federal agencies regard the "four-fifths" rule as a "rule of thumb . . . not . . . as a legal definition of adverse impact." *Adoption of Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*, *supra* note 28, Question and Answer No. 11.

Some courts have held that unless differences are statistically significant, there is no burden-shifting adverse impact. *See, e.g.*, Contreras v. City of Los Angeles, 656 F.2d 1267, 1272-73 (9th Cir. 1981), *cert. denied*, 455 U.S. 1021 (1982); Williams v. Tallahassee Motors, Inc., 607 F.2d 689, 693 (5th Cir. 1979), *cert. denied*, 449 U.S. 858 (1981). Most courts consider the "four-fifths" rule along with other evidence as measures of discriminatory effect. *See, e.g.*, Guardians Ass'n of the New York City Police Dept., Inc. v. Civil Service Comm., 630 F.2d 79, 88 (2d Cir. 1980)(adverse impact shown "[b]y any reasonable measure, including . . . the four-fifths rule . . . ."), *cert. denied*, 452 U.S. 940 (1981); Chrisholm v. United States Postal Serv., 665 F.2d 482, 495 & nn. 21-22 (4th Cir. 1981)(chi square analysis and four-fifths rule); League of Martin v. City of Milwaukee, 588 F. Supp. 1004, 1014-15 (E.D. Wis. 1984)(four-fifths rule and statistical significance); Easley v. Anheuser-Busch, Inc., 572 F. Supp. 402, 406-07 (E.D. Mo. 1983)(four-fifths rule and chi square); Burney v. City of Pawtucket, 559 F. Supp. 1089, 1099 (D.R.I. 1983)(four-fifths rule and other evidence); Berkman v. City of New York, 536 F. Supp. 177, 205 (E.D.N.Y. 1982)(four-fifths rule and binomial distribution), *aff'd*, 705 F.2d 584 (2d Cir. 1983); I.M.A.G.E. v. Bailar, 518 F. Supp. 800, 804-08 (N.D. Cal. 1981)(same); Reynolds v. Sheet Metal Workers Local 102, 498 F. Supp. 952, 965-70 (D.D.C. 1980)(four-fifths rule, chi square, and binomial distribution), *aff'd*, 702 F.2d 221 (D.C. Cir. 1981).

In Watson v. Fort Worth Bank & Trust, 108 S. Ct. 2777 (1988), a plurality noted criticism of the "four-fifths rule," and asserted that there is no "rigid mathematical formula" for establishing disparate impact. Four justices opted for a "case-by-case" approach.

45. *Uniform Guidelines, supra* note 18, §§ 3 & 6, 29 C.F.R. §§ 1607.3 & 1607.6.

46. *Id.* at § 3A, 29 C.F.R. 1607.3A. The *Uniform Guidelines* recognize three types of validation, criterion-related validity studies, content validity studies and construct validity studies, 29 C.F.R. § 1607.5A, and set forth "technical standards" for each. 29 C.F.R. § 1607.14. In Watson v. Fort Worth Bank & Trust, 108 S. Ct. 2777 (1988), a plurality said that it is not necessary "to introduce formal 'validation studies' showing that particular criteria predict actual on-the-job performance." *Id.* at 2790.

47. *Griggs*, 401 U.S. at 431; *Albemarle Paper Co.*, 422 U.S. at 425. Instead of referring to "adverse impact," in *Albemarle* the Court used the term "significantly different" selection rates. *Id.* Recently, a plurality of the Supreme Court has suggested that the plaintiff

verse impact need not prove intentional discrimination in order to make out a *prima facie* case.[48]

Before a court addresses the merits of a claim of racial discrimination in occupational testing, it must deal with one or both of two threshold issues: 1) does Title VII apply to licensing examinations? and if so, 2) does it apply to private test developers such as ETS?

1. Application of Title VII to Licensing

Whether Title VII and the *Uniform Guidelines* apply to occupational licensing is unresolved. The *Uniform Guidelines* purport to "apply to tests and other selection procedures which are used as a basis for any employment decision. Employment decisions include but are not limited to hiring, promotion, demotion, membership (for example, in a labor organization), referral, [and] retention. . . ."[49] The statement of purpose section of the *Uniform Guidelines*, however, contemplates that Title VII does apply to licensing and certification.[50] The *Uniform Guidelines*' definition of the word "user" supports this view.[51]

In determining Title VII's applicability, two EEOC decisions are instructive. EEOC Dec. No. 75-249[52] held that, inasmuch as a state insurance licensing agency itself employed more than fifteen persons and was therefore subject to the jurisdiction of Title VII, the agency was answerable to charges that it discriminated on the basis of national origin by administering its insurance examination in English only. The Commission reasoned:

> By its terms Title VII speaks not of "employees" but of "person[s] aggrieved." Throughout the Title and its legislative history Congress indicated its intent to deal with more than the

---

retains the burden throughout a "disparate impact" case. Watson v. Fort Worth Bank & Trust, 108 S. Ct. 2777 (1988).

48. Washington v. Davis, 426 U.S. 229, 247-48 (1976).

49. *Uniform Guidelines*, *supra* note 18, § 2B, 29 C.F.R. § 1607.2B.

50. The *Uniform Guidelines* state: "These guidelines incorporate a single set of principles which are designed to assist employers, labor organizations, employment agencies, and *licensing* and *certification* boards to comply with requirements of Federal law prohibiting employment practices which discriminate on grounds of race, color, religion, sex, and national origin." *Id.* at § 1B, 29 C.F.R. § 1607.1B (emphasis added).

51. The *Uniform Guidelines* provide:

> Whenever an employer, labor organization, or employment agency is required by law to restrict recruitment for any occupation to those applicants who have met licensing or certification requirements, the licensing or certifying authority to the extent it may be covered by Federal equal employment opportunity law will be considered the user with respect to those licensing or certification requirements.

Id. at § 16W, 29 C.F.R. § 1607.16W.

52. National Origin Bias Found in Use of State Insurance Licensing Exam, EEOC Decs. (CCH) ¶ 6457 (May 6, 1975).

> conventional employer-employee situation as demonstrated by the specific prohibition against discrimination by employment agencies and referral labor organizations . . . Courts have held that no employer-employee relationship need exist, only control over access to the job market and denial of such access by reference to invidious criteria.[53]

In EEOC Dec. No. 81-22,[54] the Commission found that a state police department which served as a licensing agency for private security guards was subject to Title VII.[55] In addition to relying on statutory language, the Commission also relied on the remedial purposes of Title VII, stating that "[a] proper test of the Commission's jurisdiction over the subject-matter. . . focuses on whether the Respondent acted in a manner or made a decision which adversely affected the Charging Party and allegedly deprived him of a right protected. . . by the Act."

Although the Supreme Court in *Griggs v. Duke Power Co.*[56] stated that the EEOC's construction of Title VII is entitled to "great deference," the courts have not been particularly deferential.[57] Some courts have con-

---

53. *Id.* at 4208.

54. State Licensing Agency Accountable For Job Opportunity Denial, EEOC Decs. (CCH) ¶ 6825 (May 13, 1981).

55. The EEOC explained:

> That Title VII is designed to cover more than the conventional employer-employee relationship has been recognized by numerous court cases and Commission decisions. . .Section 706(b) requires only that the charge be filed by or on behalf of "a person claiming to be aggrieved." The Charging Party is definitely aggrieved by the Respondent's disapproval of his application of a uniformed guard position if the processing of the application was conducted in a manner which discriminated against him because of his race. The disapproval of the application necessarily means that a private detective agency must by law refuse to employ or continue to employ an individual who does not obtain the Respondent's approval.

Id. at 4951 (footnote omitted).

56. 401 U.S. 424, 433-34 (1971).

57. George v. New Jersey Bd. of Veterinary Medical Examiners, 794 F.2d 113 (3d Cir. 1986); Haddock v. Bd. of Dental Examiners of California, 777 F.2d 462 (9th Cir. 1985)(state board of dental examiners is not an "employer"); EEOC v. Waterfront Comm'n, 665 F. Supp. 197 (S.D.N.Y. 1987); Lavender-Cabellero v. Dep't of Consumer Affairs, 458 F. Supp. 213, 215 (S.D.N.Y. 1978)(Title VII does not apply to a city agency in the exercise of its statutory mandate to issue licenses to process servers, finding the absence of explicit treatment of licensing by Congress to be persuasive). *Cf.* Beverly v. Douglas, 591 F. Supp. 1321 (S.D.N.Y. 1984)(Title VII does not apply to a hospital which denied the plaintiff's application for voluntary attending privileges); Darks v. City of Cincinnati, 745 F.2d 1040, 1043 (6th Cir. 1984)(Title VII does not apply to a city agency which denied the plaintiff a license to operate a dance hall).

Dictum in some cases also indicates a narrow view of the application of Title VII to licensing. *See Darks*, 745 F.2d at 1042 n.3; Tyler v. Vickery, 517 F.2d 1089, 1096 (5th Cir. 1975) (court rejected contention that bar examination was unconstitutional because it was not validated pursuant to EEOC Guidelines; the court stated that "Title VII does not apply by its

strued Title VII's definition of "employer" narrowly, holding that a government agency which licenses individuals is not an employer; thus *Woodward v. Virginia Board of Bar Examiners*[58] held that Title VII was inapplicable to professional licensing examinations. Similarly, *National Organization for Women v. Waterfront Commission*[59] held that Title VII did not apply to a commission which licensed and registered persons working on the New York waterfront. The court reasoned that Congress must have considered and rejected the application of Title VII to licensing.[60]

On the other hand, some courts have held that Title VII does reach licensing. *Sibley Memorial Hospital v. Wilson*,[61] while not directly involving licensing, articulates the more expansive view. A male private duty nurse licensed to practice in the District of Columbia alleged that the hospital discriminatorily denied him referrals to female patients on the basis of his sex, under a program whereby the hospital referred independent private duty nurses to patients. The court of appeals reversed the district court's summary judgment for the hospital, emphasizing that the objective of Title VII was "to achieve equality of employment opportunities"[62] and that this broad remedial purpose was best served by applying the Act's prohibitions to the hospital:

> Control over access to the job market may reside, depending upon the circumstances of the case, in a labor organization, an *employment agency*, or an employer as defined in Title VII; and it would appear that Congress has determined to prohibit each of these from exerting any power it may have to foreclose, on invidious grounds, access by any individual to employment opportunities otherwise available to him.[63]

---

terms, of course, because the Georgia Board of Bar Examiners is neither an 'employer,' or 'employment agency,' nor a 'labor organization' within the meaning of the statute."), *cert. denied*, 426 U.S. 940 (1976); Richardson v. McFadden, 540 F.2d 744, 747 (4th Cir. 1976) ("Appellants agree that Title VII does not apply to the bar exam by its own terms."), *rehearing en banc*, 563 F.2d 1130 (4th Cir. 1977), *cert. denied*, 435 U.S. 968 (1978).

58. 420 F. Supp. 211 (E.D. Va. 1976), *aff'd per curiam*, 598 F.2d 1345 (4th Cir. 1979).

59. 468 F. Supp. 317 (S.D.N.Y. 1979).

60. The court said:

> In its licensing role, the Commission neither pays the wages nor engages the services of persons it registers. Nor does it undertake to obtain workers for employers or jobs for workers. It is, therefore, neither an "employer" nor an "employment agency" with respect to persons desiring registration.

Id. at 320.

61. 488 F.2d 1338 (D.C. Cir. 1970).

62. *Id.* at 1340-41 (quoting Griggs v. Duke Power Co., 401 U.S. 424 (1970)).

63. *Id.* at 1341 (emphasis added). *Sibley* has been reaffirmed in Doe v. St. Joseph's Hosp. of Fort Wayne, 788 F.2d 411, 422 (7th Cir. 1986). The *Doe* court reversed the dismissal of a Title VII case brought by a physician against a hospital. Denying staff privileges, the

The principle announced in *Sibley* was applied to licensing in *Puntolillo v. New Hampshire Racing Commission.*[64] The *Puntolillo* court held that because defendants were employers under the Act and controlled plaintiff's access to his job, they were subject to the requirements of Title VII.[65]

Although the authorities are plainly divided, and the Supreme Court has yet to resolve the division, the statute's remedial purposes surely favor the broader reading of Title VII. That interpretation is particularly appro-

---

court ruled, could constitute interference with the physician's employment opportunities with her patients within the meaning of Title VII. *See also* Pardazi v. Cullman Medical Center, 838 F.2d 1155 (11th Cir. 1988); Zaklama v. Mt. Sinai Medical Center, 842 F.2d 291, 293-95 (11th Cir. 1988). *But see* Diggs v. Harris Hosp.-Methodist, Inc., 847 F.2d 270 (5th Cir. 1988), *cert. denied*, 109 S. Ct. 394 (1988).

64. 375 F. Supp. 1089 (D.N.H. 1974). A driver-trainer of harness horses alleged that defendants (the commission and a racing association which held races at a particular track) discriminatorily denied him a racing license on the basis of his national origin. The court denied motions to dismiss, holding that both defendants were "employers" within the meaning of Title VII. Admitting that the relationship was not a traditional employer-employee relationship, the court emphasized the Act's broad remedial objectives. Quoting from *Griggs*, the court stated that the purpose of the Act was to "achieve *equality of employment opportunities* and remove barriers that operated in the past." *Id.* at 1092 (emphasis by court).

65. *Id.* at 1092. Similarly in Veizaga v. National Bd. for Respiratory Therapy, 21 Fair Empl. Prac. Cas. (BNA) 246 (N.D. Ill. 1977), plaintiffs alleged that defendants, who participated in the registration or certification of respiratory therapists, violated Title VII by requiring persons seeking employment as respiratory therapists to pass certain examinations; hospitals relied on this accreditation procedure in hiring. While holding that the facts were insufficient to state a claim, the court permitted plaintiffs to amend "to allege with more specificity the manner in which NBRT controls access to employment. This should refer to some type of *licensing*, screening or channeling function which is a prerequisite to employment in various hospitals." *Id.* (emphasis added). The *Sibley* principle has been applied to civil service agencies that do not themselves employ workers. *See infra* notes 66-69. The principle also has been applied in other contexts as well. *See* Gomez v. Alexian Bros. Hosp. of San Jose, 698 F.2d 1019 (9th Cir. 1983)(applying Title VII to hospital which denied application of plaintiff's medical corporation for contract to oversee hospital's emergency room); Barone v. Hackett, 602 F. Supp. 481 (D.R.I. 1984)(applying Title VII to officials of employee benefits programs which allegedly discriminated in providing pregnancy benefits); Pao v. Holy Redeemer Hosp., 547 F. Supp. 484, 494 (E.D. Pa. 1982)(applying Title VII to hospital which denied plaintiff's application for staff privileges, where plaintiff alleged that "the defendant significantly affected or controlled his access to other employment opportunities in a discriminatory manner."); EEOC v. Sage Realty Corp., 87 F.R.D. 365, 370 (S.D.N.Y. 1980)(finding two companies to be a joint employer because the term "employer" has been construed "in a functional sense to encompass persons who are not employers in conventional terms, but who nevertheless control some aspect of an individual's compensation, terms, conditions, or privileges of employment."); Naismith v. Professional Golfers Assn., 85 F.R.D. 552, 557-61 (N.D. Ga. 1979)(applying Title VII to professional golfers' association which organized events for professional golfers); Vanguard Justice Soc'y, Inc. v. Hughes, 471 F. Supp. 670, 694-97 (D. Md. 1979)(applying Title VII to state in lawsuit concerning employment practices of city police department); Curran v. Portland Superintending School Comm., 435 F. Supp. 1063, 1073 (D. Me. 1977)(applying Title VII to city by reason of its role in appropriating funds for school system, on grounds that "the City is sufficiently involved in, and, in fact, necessary to, the total employment process that it must be considered plaintiff's employer for purposes of jurisdiction under Title VII.").

priate in light of the significance of licensing in determining access to job opportunities. In many occupations, being licensed is the *sine qua non* of being employed. As noted in *Sibley*, as a practical matter, licensing controls access to the job market. Since Congress intended to eradicate invidious restraints on access to jobs, applying Title VII to licensing furthers Congress' remedial intent.

2. Application of Title VII to Test Developers

Beyond the question of whether Title VII applies to licensing, a second issue arises as to whether, in addition to government agency users of tests, private test developers are covered by Title VII. There are at least suggestions that they are. *United States v. City of Yonkers*[66] held that the New York State Civil Service Commission was a proper defendant under Title VII where the Commission developed civil service tests used for appointments and promotions by the state and municipalities.[67] In *Vanguard Justice Society v. Hughes*,[68] the court held that a civil service commission which "exercised substantial authority and discretion in the area of testing of applicants for entry level positions with the [Police] Department" was an "employer" within the meaning of Title VII.[69] Other cases have decided, on an agency theory, that Title VII applies to civil service commissions in contexts other than testing.[70] There does not appear to be any direct holding, however, that a private test developer is an "employer" or an "employment agency" under Title VII.

B. Constitutional Protections

For the equal protection and due process provisions of the Fourteenth Amendment to apply, the conduct in question must constitute "state ac-

66. 592 F. Supp. 570 (S.D.N.Y. 1984).

67. *Id.* at 591. To the same effect is Vulcan Soc'y v. Fire Dept., 82 F.R.D. 379, 395-96 (S.D.N.Y. 1979)(state civil commission is subject to Title VII to the extent it prepared the test used in connection with municipal hiring).

68. 471 F. Supp. 670, 696-97 (D. Md. 1979).

69. *Id.* at 696. Other cases have held that Title VII applies to civil service commissions in contexts other than testing. *See, e.g.*, Williams v. City of Montgomery, 742 F.2d 586, 588-89 (11th Cir. 1984)(Title VII applies to city personnel board since it exercises functions "traditionally exercised by an employer"), *cert. denied*, 470 U.S. 1053 (1985); Cannon v. State of Delaware, 523 F. Supp. 341, 344 (D. Del. 1981)(Title VII applies to state personnel commission).

70. A leading text states that "[W]here the employer had delegated control of some of the employer's traditional rights, such as hiring or firing, to a third party, the third party has been found to be an "employer" by virtue of the agency relationship." B. SCHLEI & P. GROSSMAN, EMPLOYMENT DISCRIMINATION LAW at 1002 (2d ed. 1983) (footnote omitted) [hereinafter B. SCHLEI & P. GROSSMAN].

tion," since private acts are not proscribed.[71] While a state licensing agency is clearly engaged in state action,[72] it is uncertain whether a private test developer of a licensing test is also subject to constitutional scrutiny.

1. State Action

The Supreme Court has repeatedly stated that the issue of state action is fact-dependent,[73] and in analyzing the relevant facts it has suggested several approaches.[74] Of particular significance to occupational testing is the "public function" analysis. In *Evans v. Newton*[75] the Court held that the actions of private trustees appointed to administer a racially segregated park constituted state action because "the tradition of municipal control had become firmly established."[76] More recently the Court explained that to be treated as a state actor a nominally private party must exercise a power which is "traditionally the exclusive prerogative of the State."[77] A wide variety of functions has been characterized as traditionally governmental, including law enforcement,[78] fire protection,[79] evaluation of judicial candidates,[80] and administering a local library.[81]

---

71. Rendell-Baker v. Kohn, 457 U.S. 830, 837 (1982); Shelly v. Kraemer, 334 U.S. 1, 13 (1948).

72. *See generally* 51 AM. JUR. 2D *Licenses & Permits* § 14 (2d ed. 1970).

73. *See, e.g.*, Burton v. Wilmington Parking Auth., 365 U.S. 715, 722 (1961).

74. *See, e.g.*, Shelley v. Kraemer, 334 U.S. 1 (1948); Burton v. Wilmington Parking Auth., 365 U.S. 715 (1961); Evans v. Newton, 382 U.S. 296 (1966); Jackson v. Metropolitan Edison Co., 419 U.S. 345 (1974); Flagg Bros., Inc. v. Brooks, 436 U.S. 149 (1978); Rendell-Baker v. Kohn, 457 U.S. 830 (1982); Blum v. Yaretsky, 457 U.S. 991 (1982); Lugar v. Edmondson Oil Co., 457 U.S. 922 (1982).

75. 382 U.S. 296 (1966).

76. *Id.* at 301.

77. Jackson v. Metropolitan Edison Co., 419 U.S. 345, 353 (1974)(the function must be exclusively a state function); Rendell-Baker v. Kohn, 457 U.S. 830, 842 (1982)(although special education for "maladjusted high school students is a public function," it is not "the exclusive province of the state"); Blum v. Yaretsky, 457 U.S. 991, 1012 (1982)("decisions made in day-to-day administration of a nursing home" are not "the kind of decisions traditionally and exclusively made by the sovereign for and on behalf of the public."); Watkins v. Reed, 557 F. Supp. 278, 281-82 (E.D. Ky. 1983), *aff'd*, 734 F.2d 17 (6th Cir. 1984); Avallone v. Wilmington Medical Center, Inc., 553 F. Supp. 931, 934 (D. Del. 1982); Spencer v. General Tel. Co., 551 F. Supp. 896, 898 (M.D. Pa. 1982). *See also* Hodges v. Metts, 676 F.2d 1133, 1137 (6th Cir. 1982); Newsom v. Vanderbilt Univ., 653 F.2d 1100, 1114 (6th Cir. 1981); Musso v. Suriano, 586 F.2d 59, 63 (7th Cir. 1978), *cert. denied*, 440 U.S. 971 (1979); Lowell v. Wantz, 85 F.R.D. 286, 288 (E.D. Pa. 1980), *aff'd*, 636 F.2d 1209 (3d Cir. 1980). It is thus not enough that a party's activity be affected with a public interest. *See* Gerena v. Puerto Rico Legal Servs., Inc., 697 F.2d 447, 451 (1st Cir. 1983); Nguyen v. United States Catholic Conference, 548 F. Supp. 1333, 1344 (W.D. Pa. 1982), *aff'd*, 719 F.2d 52 (3d Cir. 1983). *See also* Bloomer Shippers Ass'n v. Illinois Central Gulf R.R., 655 F.2d 772, 776 (7th Cir. 1981).

78. Henderson v. Fisher, 631 F.2d 1115, 1118 (3d Cir. 1980).

79. Janusaitis v. Middlebury Volunteer Fire Dept., 607 F.2d 17, 22-25 (2d Cir. 1979).

80. Rouse v. Judges of Circuit Ct. of Cook County, 609 F. Supp. 243, 247 (N.D. Ill.

Occupational licensing is traditionally an exclusive state function.[82] Where a private firm develops a test which is an essential part of the licensing process, its conduct logically should constitute state action subject to due process and equal protection limitations. In the *Golden Rule* case[83] the Illinois Appellate Court reversed the lower court's dismissal, holding that, as alleged, the "[Illinois Insurance] Director and ETS have been ineluctably intertwined. . . ."[84] The court noted the substantial nature of the tasks performed by ETS in the licensing process.[85] Furthermore, the court found that passing the ETS-prepared test was the "*sine qua non* of the ability to engage in the insurance agent or broker business in the very first instance."[86] ETS has also been held to be subject to the due process clause

---

1985).

81. Chalfant v. Wilmington Inst., 574 F.2d 739 (3d Cir. 1978). Other functions have been held not to be public. *See, e.g.*, Jackson v. Metropolitan Edison Co., 419 U.S. 345 (1974); Flagg Bros., Inc. v. Brooks, 436 U.S. 149 (1978). At one time, the Supreme Court appeared willing to embrace a broader view of state action, stating that "[i]t is enough that he [the alleged wrongdoer] is a willful participant in joint activity with the State or its agents." Adickes v. S.H. Kress & Co., 398 U.S. 144, 152 (1970)(state action by private restaurant if it is shown that restaurant employee reached "an understanding" with police officer to refuse to serve plaintiff or to cause her subsequent arrest)(quoting United States v. Price, 383 U.S. 787, 794 (1966)).

See Fitzgerald v. Mountain Laurel Racing, Inc., 607 F.2d 589 (3d Cir. 1979), *cert. denied*, 446 U.S. 956 (1980), where the court first concluded that there was not a "symbiotic relationship," as defined in Burton v. Wilmington Parking Auth., 365 U.S. 715 (1961), between the state and a private racing association so that "every act of [the association] is an act of the State." The court then held, however, that there was state action in light of the "nexus" between the association's particular actions and the state, within the meaning of Jackson v. Metropolitan Edison Co., 419 U.S. 345 (1974). 607 F.2d at 597-600. *See also* Kissinger v. New York City Transit Auth., 274 F. Supp. 438, 441 (S.D.N.Y. 1967) (advertising firm under contract with transit authority subject to first and fourteenth amendments).

82. *See generally* 51 AM. JUR. 2D *Licenses & Permits* § 14 (2d ed. 1970).

83. Golden Rule Life Ins. Co. v. Mathias, 86 Ill. App. 3d 323, 408 N.E.2d 310 (1980).

84. *Id.* at 330, 408 N.E.2d at 316.

85. *Id.* The court noted that under the contract between ETS and the Director, ETS performed the following functions: (1) developed the examination in accordance with specifications developed by ETS; (2) retained complete control over the examination, including ownership of the copyright, and physical control over printed copies of the examination; (3) printed the examination, established testing centers and provided personnel to administer the examination; (4) processed all applications of those who desired to take the examination and become licensed; (5) graded all examinations and determined who has passed; and (6) printed the State licenses for the Director.

86. *Id.* at 331, 408 N.E.2d at 316. The court distinguished another case, Stewart v. Hannon, 469 F. Supp. 1142 (N.D. Ill. 1979), where passing an examination used to determine eligibility did not assure appointment to the position of school principal. *Id.*

While ETS's involvement in the licensing process in *Golden Rule* was more substantial than its involvement with the selection of principals in *Stewart* by virtue of the relative significance of the examination in the two processes, the court in *Stewart* misapplied the state action doctrine because the test at issue there was more than a significant part of the process—it was *the* principal means of selection. Failure to pass the test was a conclusive bar for those seeking to become principals. *Cf.* Connecticut v. Teal, 457 U.S. 440 (1982) (employer liable under

where it acted as the agent of the Commonwealth of Pennsylvania in developing and administering its multistate real estate licensing examination.[87]

The Supreme Court's most recent state action decision does not resolve the issue. In *National Collegiate Athletic Ass'n v. Tarkanian*,[88] the Supreme Court held that the NCAA's actions leading to a state university's suspension of an intercollegiate athletic coach did not constitute state action. The majority identified two analytically distinct contexts in which the "state action" issue arises. "In the typical case," said the Court, ". . . . a private party has taken the decisive step that caused the harm to the plaintiff . . . "[89] The question there is "whether the State was sufficiently involved to treat that decisive conduct as state action."[90] According to the majority, the case before it presented the other, non-typical, analytical context in which the "state action" issue arises: where the private party participated in the challenged conduct, but the state itself took the "decisive step."[91] The majority concluded that the NCAA was not controlled by a state and that the state university's decision to adopt the NCAA's standards or procedures did not render the NCAA a state actor.[92] Finally, while acknowledging that a private party may become a state actor by virtue of a state's delegation of state authority,[93] the majority found that the state had not delegated authority to the NCAA; rather the NCAA and the state university " . . . acted much more like adversaries than like partners. . . ."[94]

Title VII for racial discrimination with respect to written test even where other factors in promotion process compensate for discriminatory impact of test). Having been delegated substantial responsibility for the test, ETS's conduct should have been deemed state action. Immediately after plaintiffs filed their notice of appeal, defendants adopted a formal resolution not to use the results of the examination, thereby rendering the appeal moot. *See* Stewart v. Hannon, 675 F.2d 846, 848 (7th Cir. 1982).

87. Martin v. Educational Testing Serv., Inc., 179 N.J. Super. 317, 431 A.2d 868 (1981). However, educational admissions tests have been held to be beyond the reach of constitutional limitations. *See, e.g.*, Johnson v. Educational Testing Serv., 754 F.2d 20 (1st Cir. 1985), *cert. denied*, 472 U.S. 1029 (1985). The *Johnson* court distinguished the *Golden Rule* and *Martin* cases, noting that its ruling "is not to say that ETS can never be a state actor engaging in state action." 754 F.2d at 25 n.2. On the other hand, discrimination in connection with state scholarship awards based solely on SAT scores is constitutionally actionable. *See* Sharif v. N.Y. State Educ. Dept., No. 88-CIV-8435 (JMW)(S.D.N.Y. Feb 3, 1989).

88. 109 S. Ct. 454 (1988).

89. *Id.* at 462.

90. *Id.* The majority noted three situations where such "decisive conduct" may occur: (1) if the State creates the legal framework governing the conduct; (2) if the State delegates its authority to the private actor; or (3) sometimes if the State knowingly accepts the benefits derived from unconstitutional behavior. *Id.* at 463-65.

91. *Id.* at 464.

92. *Id.* at 462-65.

93. *Id.* at 464.

94. *Id.* The majority also noted that the NCAA "enjoyed no governmental powers to facilitate its investigation." *Id.* Further, a private party does not become a state actor even where a " . . . private monopolist can impose its will on a state agency by a threatened refusal

The four dissenting Justices concluded that it was sufficient that the NCAA and the state university "acted jointly" in suspending the coach.[95] They contended that, in previous decisions, the fact that the state carried out the final or decisive act had not stood in the way of a finding that a private entity was a state actor.[96]

In the typical licensing case, the extent of the powers delegated to a private test developer surely satisfies the "joint actor" standard of the dissent in *NCAA v. Tarkanian.* Moreover, the extent of the delegation may be sufficient to satisfy the more demanding "decisive step" standard, particularly since the majority noted with approval an earlier decision holding that a private physician who contracted with a state prison to provide care for inmates was in fact a state actor.[97] Significantly, unlike the "adversarial" relationship stressed by the majority in *NCAA v. Tarkanian*, the state and the test developer are true partners in the licensing process. In any event, the different legal standards articulated by the Court and the fact-dependent nature of the "state action" analysis lead to considerable uncertainty as to the extent to which constitutional protections apply to private developers of licensing examinations.

2. Equal Protection

The Supreme Court requires that a plaintiff alleging an equal protection claim prove intentional discrimination.[98] A showing of discriminatory effect which would suffice under Title VII is insufficient to prove a violation of the equal protection clause of the Fourteenth Amendment.[99] Nevertheless, "[d]isproportional impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution."[100] The Court explained that "[n]ecessarily, an invidious discriminatory purpose may often be inferred from the totality of the rele-

---

to deal with it . . . . " *Id.* at 465. Finally, the majority found that the NCAA's conduct was not "fairly attributable to the state" because the state opposed the NCAA's sanctions; at most the state university "conducted its athletic program under color of the policies adopted by the NCAA, rather than that those policies were developed and enforced under color of Nevada law." *Id.* at 465.

95. *Id.* at 466. The dissent noted that the state university suspended the coach for violation of NCAA rules pursuant to its agreement with the NCAA. *Id.* Further, the university agreed that the NCAA would conduct the hearings concerning the coach's violation of NCAA rules. *Id.* at 466-67. Finally, the university agreed that it would be bound by the NCAA's findings. *Id.* at 467.

96. *Id.* at 466 (citing Adickes v. S.H. Kress & Co., 398 U.S. 144 (1970) and Dennis v. Sparks, 449 U.S. 24 (1980)).

97. *Id.* at 462 (citing West v. Atkins, 108 S. Ct. 2250 (1988)).

98. Washington v. Davis, 426 U.S. 229, 242 (1976).

99. *See, e.g.*, Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

100. 426 U.S. at 242.

vant facts, including the fact, if it is true, that the law bears more heavily on one race than another."[101]

The Court has enumerated factors to be considered in establishing discriminatory intent.[102] As the Court has stated: "[d]etermining whether invidious discriminatory purpose was a motivating factor demands a sensitive inquiry into such circumstantial and direct evidence of intent as may be available."[103] Discriminatory impact "may provide an important starting point"[104] and in some cases may be so substantial that, if "unexplainable on grounds other than race," it may suffice. Also relevant are "[t]he historical background," "the specific sequence of events leading up to the challenged decision," "[d]epartures from the normal procedural sequence," "substantive departures," and "[t]he legislative or administrative history."[105]

In *Personnel Administrator of Massachusetts v. Feeney*,[106] the Court appeared to articulate a stricter view of the proof needed to invalidate a state statute having a discriminatory effect. There the Court upheld a veterans' preference statute which clearly had a discriminatory impact on women:

> "Discriminatory purpose," however, implies more than intent as volition or intent as awareness of consequences. It implies that the decisionmaker, in this case a state legislature, selected or reaffirmed a particular course of action at least in part "because of," not merely "in spite of," its adverse effects upon an identifiable group.[107]

But the Court did not reject discriminatory impact as an element of proving discriminatory intent. Quoting *Village of Arlington Heights v. Metropolitan Housing Development*, the Court acknowledged not only that "impact provides an 'important starting point,' "[108] but also that "there are cases in which impact alone can unmask an invidious classification."[109] The Court further explained that "[i]f the impact of this statute could not be plausibly explained on a neutral ground, impact itself would signal that the real classification made by the law was in fact not neutral."[110] Further, "[p]roof of discriminatory impact must necessarily usually rely on objective factors,

---

101. *Id.*
102. Village of Arlington Heights v. Metropolitan Hous. Dev. Corp., 429 U.S. 252, 266 (1977).
103. *Id.* at 266.
104. *Id.*
105. *Id.* at 267-68.
106. 442 U.S. 256 (1979).
107. *Id.* at 279 (footnotes omitted).
108. *Id.* at 274 (quoting *Arlington Heights*, 429 U.S. at 266).
109. *Feeney*, 442 U.S. at 275.
110. *Id.*

several of which were outlined in *Village of Arlington Heights* . . . . The inquiry is practical. What a legislature or any official entity is 'up to' may be plain from the results its actions achieve, or the results they avoid."[111] In a testing case, discriminatory impact is a major element of a plaintiff's proof. The Supreme Court certainly has suggested that extreme disparate impact may by itself support a finding of discriminatory intent.[112]

The Sixth Circuit has held that "[s]tatistical evidence of racially disparate impact of employment practices alone may establish a statutory violation. . . . Under some circumstances, such evidence may also demonstrate a constitutional violation."[113] Other federal courts have also recognized the significance of extreme disparate impact in proving intent.[114]

Even if discriminatory impact by itself is not sufficient to establish the requisite intent, the foreseeability of such impact is probative of intent.[115]

---

111. *Id.* at 279 n.24. *Feeney* can also be read as establishing a presumption of discriminatory intent upon a showing of discriminatory effect. The court stated:

> This is not to say that the inevitability or foreseeability of consequences of a neutral rule has no bearing upon the existence of discriminatory intent. Certainly, when the adverse consequences of a law upon an identifiable group are as inevitable as the gender-based consequences of [the veterans' preference statute], a strong inference that the adverse effects were desired can reasonably be drawn.

442 U.S. at 279 n.25. *See also* Columbus Bd. of Educ. v. Penick, 443 U.S. 449 (1979). Thus, the Court in *Feeney* in effect held that the presumption of intent had been rebutted by proof of a non-invidious purpose, and "the inference simply fail[ed] to ripen into proof." 442 U.S. at 279 n.25.

112. Arlington Heights v. Metropolitan Hous. Dev. Corp., 429 U.S. 252 at 266 (1977). *See also* Batson v. Kentucky, 476 U.S. 79, 93 (1986); Casteneda v. Partida, 430 U.S. 482, 494 n.13 (1977)("If a disparity is sufficiently large, then it is unlikely that it is due solely to chance or accident, and, in the absence of evidence to the contrary, one must conclude that racial or other class-related factors entered into the selection process."); Turner v. Fouche, 396 U.S. 346, 359 (1970)(comparing number of Blacks on grand jury list with number to be expected from random selection).

113. Detroit Police Officers' Ass'n v. Young, 608 F.2d 671, 686 (6th Cir. 1979), *cert. denied*, 452 U.S. 938 (1981).

114. Dowdell v. City of Apopka, 698 F.2d 1181, 1186 (11th Cir. 1983); United States v. Texas Educ. Agency, 600 F.2d 518, 528 (5th Cir. 1979); Louisville Black Police Officers Org., Inc. v. City of Louisville, 511 F. Supp. 825, 832 (W.D. Ky. 1979), *aff'd*, 700 F.2d 268 (6th Cir. 1983). *Cf.* Payne v. Travenol Labs., Inc., 673 F.2d 798, 817 (5th Cir. 1982)(in Title VII disparate treatment case, *prima facie* case of intent may be established, under *Arlington Heights*, "solely with statistics if they show a sufficiently great disparity between the employer's treatment of blacks and of whites — a disparate result. In such circumstances, statistics alone justify an inference of discriminatory motive."), *cert. denied*, 459 U.S. 1038 (1982).

115. Detroit Police Officers' Ass'n v. Young, 608 F.2d 671, 693 (6th Cir. 1979)(discriminatory intent may be established by any evidence—including "actions having foreseeable and anticipated disparate impact . . . which logically supports an inference that the state action was characterized by invidious purposes."), *cert. denied*, 452 U.S. 938 (1981); United States v. Texas Educ. Agency, 579 F.2d 910, 913 n.5 (5th Cir. 1978), *cert. denied*, 443 U.S. 915 (1979); United States v. Texas Educ. Agency, 564 F.2d 162, 165-78 (5th Cir. 1977); Diaz v. San Jose School Dist., 733 F.2d 660, 663 (9th Cir. 1984)(*en banc*), *cert. denied*, 471 U.S.

Disparate impact is of particular significance where it is the inevitable and foreseeable consequence of a defendant's actions. The Supreme Court made this point in *Columbus Board of Education v. Penick*,[116] less than a month after *Feeney*.[117]

The fact that a defendant is aware of the discriminatory impact of its actions but fails to take steps to investigate or remedy the problem is probative of discriminatory intent.[118] The Illinois Appellate Court in *Golden Rule*

1065 (1985); Dowdell v. City of Apopka, 698 F.2d 1181, 1186 (11th Cir. 1983); Alexander v. Youngstown Bd. of Educ., 675 F.2d 787, 792 (6th Cir. 1982); Hoots v. Commonwealth of Pennsylvania, 672 F.2d 1107, 1114, 1116 (3d Cir.), *cert. denied*, 459 U.S. 824 (1982); United States v. School Dist. of Omaha, 565 F.2d 127 (8th Cir. 1977), *cert. denied*, 434 U.S. 1064 (1978); Larry P. v. Riles, 495 F. Supp. 926, 976 (N.D. Cal. 1979)(characterizing the reliance on foreseeability as consistent with "a majority of federal appellate courts."), *modified on appeal*, 793 F.2d 969, 984 (9th Cir. 1984)("pervasiveness of discriminatory effect" cannot "without more, be equated with the discriminatory intent required by *Washington v. Davis*"). *See also* Arthur v. Nyquist, 573 F.2d 134, 141-43 (2d Cir.), *cert. denied*, 439 U.S. 860 (1978); United States v. Board of School Commr's of City of Indianapolis, 573 F.2d 400, 413 (7th Cir.), *cert. denied*, 434 U.S. 997 (1977).

This emphasis on the probative value of disparate impact where it was foreseeable is appropriate in view of the principle that actions to enforce the constitutional right of equal protection under 42 U.S.C. § 1983 "should be read against the background of tort liability that makes a man responsible for the natural consequence of his action." Monroe v. Pape, 365 U.S. 167, 187 (1961). *See* Washington v. Davis, 426 U.S. 229, 253 (1976)(Stevens, J. concurring); United States v. Texas Educ. Agency, 564 F.2d 162, 167-68 (5th Cir. 1977); United States v. Texas Educ. Agency, 579 F.2d 910, 913 (5th Cir. 1978)(*Monroe v. Pape* rule "is the most reliable [judicial mechanism] for the objective determination of intent.").

116. 443 U.S. 449 (1979).

117. In *Penick*, the Court said that:

> . . . the District Court correctly noted that actions having foreseeable and anticipated disparate impact are relevant evidence to prove the ultimate fact, forbidden purpose. . . Adherence to a particular policy or practice, with full knowledge of the predictable effects of such adherence upon racial imbalance in a school system is one factor among many others which may be considered by a court in determining whether an inference of segregative intent should be drawn. The District Court thus stayed well within the requirements of *Washington v. Davis* and *Arlington Heights*.

Id. at 464-65. While the Supreme Court stated, in Dayton Bd. of Educ. v. Brinkman, 443 U.S. 526, 536-37 n.9 (1979), that, as a general proposition, the foreseeability of segregative effects does not by itself make out a *prima facie* case of intentional discrimination or "routinely" shift the burden of proof to a defendant, the Court did not reject the application of such a presumption *per se*; the presumption may still be applied in a proper case. *See, e.g.*, Larry P. v. Riles, 495 F. Supp. 926, 979 (N.D. Cal. 1979), *modified on appeal*, 793 F.2d 969 (9th Cir. 1984).

118. *Penick*, 443 U.S. at 465 (1979)(quoting district court). *See also* Rogers v. Lodge, 458 U.S. 613, 626 (1982); Richardson v. Pennsylvania Dept. of Health, 561 F.2d 489, 492 (3d Cir. 1977); Dickerson v. United States Steel Corp., 472 F. Supp. 1304, 1352-53 (E.D. Pa. 1978) *vacated and remanded by* Worthy v. United States Steel Corp., 616 F.2d 698 (3d Cir. 1980); Delgado v. McTighe, 442 F. Supp. 725, 727-28 (E.D. Pa. 1977); Arnold v. Ray, 21 Fair Empl. Prac. Cas. (BNA) 793 (N.D. Ohio 1979); Price v. Denison Indep. School Dist., 694 F.2d 334, 371 (5th Cir. 1982); Hoots v. Commonwealth of Pennsylvania, 672 F.2d 1107, 1118 (3d Cir.), *cert. denied*, 459 U.S. 824 (1982); United States v. Texas Educ. Agency, 564 F.2d 162 (5th Cir. 1977). *Cf.* Diaz v. San Jose Unified School Dist., 733 F.2d 660, 663 n.1

upheld a claim under the equal protection clause that ETS and the Illinois Insurance Director engaged in intentional discrimination by failing to act in the face of test results which clearly showed significant disparate racial impact.[119] As a matter of policy and logic, a failure to act in the face of known discriminatory effect should be strongly probative of intentional discrimination. Where the racial disparities in test results are statistically significant, they are not explainable by chance.[120] For a defendant to allow such disparities to continue without ascertaining the reason constitutes at the very least reckless disregard for the consequences. In the absence of a compelling showing to the contrary, inaction under such circumstances is explainable only as willful omission. To find a constitutional violation under such circumstances is entirely consistent with *Washington v. Davis* and its progeny. It is uncertain, however, whether the Supreme Court is likely to go this far.

3. Due Process

Alternatively, the due process clause may serve as the basis for a constitutional claim even in the absence of class-based discrimination. The concept of freedom to engage in an occupation has roots in early American history.[121] The Supreme Court has noted that the concept of liberty "denotes not merely freedom from bodily restraint but also the right of the individual to contract, to engage in any of the common occupations of life, to acquire useful knowledge . . . and generally to enjoy those privileges long recognized . . . as essential to the orderly pursuit of happiness of free men."[122] The Court also has said that "the right to work for a living in the common occupations of the community is of the very essence of the personal freedom and opportunity that it was the purpose of the [Fourteenth] Amendment to secure."[123]

As early as 1923, the Court announced the applicable due process standard:

(9th Cir. 1984)(*en banc*)("[w]here, as here, the adverse consequences are clearly identified and repeatedly articulated to the decisionmaking body, the inevitability of the adverse effects provides a strong inference of illegitimate intent."), *cert. denied*, 471 U.S. 1065 (1985). *But cf.* United States v. City of Chicago, 549 F.2d 415, 435 (7th Cir. 1977)("that the [defendant] must have been aware" of racially disparate impact was not enough to show purposeful discrimination), *cert. denied sub nom.*, Arado v. United States, 434 U.S. 875 (1977).

119. Golden Rule Life Ins. Co. v. Mathias, 86 Ill. App. 3d 323, 408 N.E.2d 310 (1980).

120. *See* B. SCHLEI & P. GROSSMAN, *supra* note 70 at 98-99.

121. *See, e.g.*, J. MADISON, *Essay on Property & Liberty* (1792) in THE COMPLETE MADISON—HIS BASIC WRITINGS 268 (S. Padover ed. 1953).

122. Meyer v. Nebraska, 262 U.S. 390, 399 (1923).

123. Traux v. Raich, 239 U.S. 33, 41 (1915). In Hampton v. Mow Sun Wong, 426 U.S. 88, 102 (1976), the Court quoted with approval the passage quoted from *Traux*, 239 U.S. at 41.

> If [a state] purported to confer arbitrary discretion to withhold a license, or to impose conditions which have no relation to the applicant's qualifications to practice [a profession] the statute would, of course, violate the due process clause of the Fourteenth Amendment.[124]

The Court reaffirmed the application of the due process clause to occupational licensing in 1957, holding that the New York bar examiners deprived an applicant of due process by refusing to grant him a license to practice law.[125] The Court stated that "a person cannot be prevented from practicing except for valid reasons."[126] To be constitutional, "any qualification must have a rational connection with the applicant's fitness or capacity to practice law."[127]

Lower courts have applied this standard to occupational licensing.[128] In particular, *Martin v. Educational Testing Service*[129] stated that there is "a substantive right to be tested fairly and accurately."[130] Accordingly, if it can be shown that a licensing examination arbitrarily excludes qualified persons from an occupation, the due process clause provides a remedy. The challenge, however, is to establish that a test is sufficiently arbitrary to violate due process.

The *Golden Rule* case is illustrative. The case arose as a consequence of the multistate insurance agents licensing examination developed by ETS with the sponsorship of the National Association of Insurance Commission-

---

124. Douglas v. Noble, 261 U.S. 165, 168 (1923).

125. Schware v. Board, of Examiners, 353 U.S. 232, 238-39 (1957).

126. *Id.* at 239 n.5 (1957).

127. *Id.* at 239.

128. *See, e.g.*, Richardson v. McFadden, 540 F.2d 744, 750-51 (4th Cir. 1976)(bar examiner "acted arbitrarily and capriciously in violation of both Due Process and Equal Protection Clauses" in grading examinations and in deciding who passed in "borderline" cases)(insufficient proof of due process violation), *decision vacated on rehearing en banc*, 563 F.2d 1130 (4th Cir. 1977), *cert. denied*, 435 U.S. 968 (1978); Brown v. Supreme Ct. of Nevada, 476 F. Supp. 86 (D. Nev. 1979), *rev'd on other grounds*, Brown v. Board of Bar Examiners, 623 F.2d 605 (9th Cir. 1980). *But see* Bigby v. City of Chicago, 766 F.2d 1053, 1058 (7th Cir. 1985)(stating in dictum that, in the absence of claims of racial discrimination, a promotional examination could be deemed to violate the due process clause only if it "shock[s] the conscience"), *cert. denied sub nom.* Thoele v. City of Chicago, 474 U.S. 1056 (1986).

129. 179 N.J. Super. 317, 431 A.2d 868, 872 (1981).

130. *Id.* at 321, 431 A.2d at 872. There is also a related procedural due process right, which requires that a state or government agency may not impose an examination requirement as a condition for receiving an entitlement (such as a high school diploma or a post-graduate degree) without notice sufficient to give affected persons time to prepare for the examination. *See, e.g.*, Debra P. v. Turlington, 644 F.2d 397, 403-04 (5th Cir. 1981)(inadequate notice of requirement for receipt of diploma to pass exit examination); Brookhart v. Illinois State Bd. of Educ., 697 F.2d 179 (7th Cir. 1983); Mahavongsanan v. Hall, 529 F.2d 448 (5th Cir. 1976) (adequate notice of new comprehensive examination requirement for receipt of master's degree).

ers (NAIC). Illinois was the first state to offer the test, in late 1975. In mid-1976, Golden Rule Insurance Company and five individuals who had failed the examination brought suit against ETS and the Director of the Illinois Insurance Department. In addition to their claims of racial discrimination, plaintiffs alleged that ETS and the Illinois Insurance Director violated the due process clause with respect to the insurance agents licensing examination. They claimed that the examination covered subject areas inappropriate to an entry-level examination, was complex and confusing in form and structure and contained obscure and highly technical questions, knowledge of which was inappropriate to an entry-level examination. The test allegedly contained many questions subject to different interpretations and different answers by individuals experienced and competent as insurance agents and brokers. Additionally, the exam allegedly tested levels of cognition of subject matter substantially and rationally unrelated to a determination of an applicant's competency as an insurance agent or broker. The plaintiffs alleged that the test was given without any job validation to determine whether in fact it appropriately measured competency to engage in the business of an insurance agent or broker, and was not fairly designed to measure an applicant's competency. Instead, the test served as a method of artificially limiting and controlling the number of individuals entering the business of insurance agent or broker without regard for competency.[131]

Although the Illinois Appellate Court expressed doubt that these allegations could be sustained at trial,[132] it held that they stated a claim under the due process clauses of the federal and state constitutions.[133] Apparently then, the due process clause of the Fourteenth Amendment provides some protection against arbitrary licensing examinations, even in the absence of claims of racial or other discrimination.

131. Golden Rule Life Ins. Co. v. Mathias, 86 Ill. App. 3d 323, 326-27, 408 N.E.2d 310, 312-14 (1980).

132. *Id.* at 333, 408 N.E.2d at 318.

133. *Id.* at 333, 408 N.E.2d at 319.

IV. THE SIGNIFICANCE OF RACIAL DATA

Both the *Uniform Guidelines*[134] and the APA *Standards*[135] require the collection and analysis of test data by race,[136] as did the earlier 1970 EEOC Guidelines on Employee Selection Procedures.[137] To determine whether a test has an adverse impact on any protected racial group, racial statistics are obviously required.

When Illinois became the first state to adopt the ETS insurance licensing examinations in 1975, it had the assurance of the National Association of Insurance Commissioners, the initial sponsor of the ETS program,[138] that the tests were "designed to be free from objection by the EEOC."[139] ETS has professed for years that it "will adhere to the appropriate professional standards" such as the APA *Standards*.[140] The APA's 1985 revised Standard 3.10 provides:

> When previous research indicates the need for studies of item or test performance differences for a particular kind of test for members of age, ethnic, cultural, and gender groups in the population of test takers, such studies should be conducted as soon as is feasible. Such research should be designed to detect and eliminate aspects of test design, content, or format that might bias test scores for particular groups.[141]

---

134. *Uniform Guidelines, supra* note 18.

135. *See supra* note 17. Courts have looked to the APA *Standards* in testing cases. Washington v. Davis, 426 U.S. 229, 247 n.13 (1976)(noting that APA *Standards* have been relied upon by the EEOC in its guidelines and "have been judicially noted in cases where validation of employment tests has been in issue [citing cases]"); Douglas v. Hampton, 512 F.2d 976, 984 n.59 (D.C. Cir. 1975)(APA *Standards* are "[t]he universally recognized professional authority" with regard to the techniques for establishing test validity); Walls v. Mississippi Dept. of Public Welfare, 542 F. Supp. 281, 313-14 (N.D. Miss. 1982)("[t]est developers are also required to follow standards established by the [APA] . . . . ").

136. *Uniform Guidelines, supra* note 18, § 4, 29 C.F.R. § 1607.4.

137. *Guidelines on Employee Selection Process* 35 Fed. Reg. 12,333-34 (1970).

138. The NAIC sponsored ETS's multi-state insurance licensing testing program from that program's inception in 1975. *See* 1976 Vol. I NAIC *Proceedings* at 280. In 1984, the NAIC withdrew its sponsorship. *See* 1985 Vol. I NAIC *Proceedings* at 152; ETS *1986 MILP Advisory Board Meeting* at 1 (1986).

139. 1975 Vol. II NAIC *Proceedings* at 252. Whether the tests were in fact developed in accordance with the EEOC *Guidelines* was an issue which had not been resolved by the court at the time *Golden Rule* was settled.

140. ETS *Standards for Quality and Fairness* 10 (1983). A version of ETS *Standards* published in 1981 had the same provision. ETS, *Standards For Quality and Fairness* 12 (1981). [*hereinafter* ETS *Standards*] Specifically with respect to its occupational examinations, ETS has said that "COPA tests meet the highest professional standards for validity, job-relatedness, reliability, and equating [comparability across forms]." *The Center for Occupational & Professional Assessment, Licensing, Certification & Assessment* 4 (ETS 1983).

141. APA *Standards supra* note 17, § 3.10. This standard is denoted "Conditional." A

The APA's comment on Standard 3.10 suggests that, while such studies may not be feasible prior to the first use of a test, operational use provides the occasion for assembling and studying the data to determine if there is bias.[142]

In October 1981, ETS's trustees formally adopted a set of internal guidelines, which previously had been in use at ETS.[143] One guideline provided that ETS should engage in research to maintain the quality of operational programs, including "studies to determine the sources of significant differential performance of sex, ethnic, handicapped, and other relevant subgroups on ETS tests."[144] Another guideline highlighted the importance of descriptive statistics "in order to monitor the participation and performance of males and females drawn from diverse backgrounds, interests and experience (e.g., major ethnic group handicapped status and other relevant subgroups of the population of interest.)"[145] A third guideline, relating to "Test Validity," provided that when appropriate and feasible "the validity of a test should be investigated separately for subsamples of the test-taking population. . . . "[146] A later edition of ETS's *Standards* retained the substance of these guidelines.[147]

It is therefore most troublesome that, from the introduction in October 1975 of the ETS-prepared insurance licensing examinations until the 1984 *Golden Rule* settlement, ETS did not collect or analyze by race any Illinois test statistics.[148] Even more disturbing is the failure of the Illinois Department of Insurance to insist that ETS do so. Nevertheless, the information that was available suggested that there was a problem. Some 1977-78 statistics compiled by ETS from comparable insurance examinations in Wis-

---

conditional standard "should be considered primary [i.e., "should be met by all tests before their operational use and in all test uses, unless a sound professional reason is available to show why it is not necessary, or technically feasible, to do so in a particular case"] for some situations and secondary [i.e., "desirable as goals but are likely to go beyond reasonable expectations in many situations"] for others." APA *Standards*, *supra* note 17, Introduction at 3 (1985 ed.). "[I]f the use of a test is likely to have serious consequences for test takers, especially if a large number of people may be affected, conditional standards assume increased importance." *Id.* Although Standard 3.10 only requires studies of item or test performance differences "when previous research indicates the need for [such studies]," it is at least arguable that, given the widely observed phenomenon of disparities in item and test performance as between racial and ethnic groups, studies of "performance differences for a particular kind of test" are required by Standard 3.10. *Id.*

142. APA *Standards*, *supra* note 17, at 27.

143. ETS *Standards*, *supra* note 140, at iii, 21 (1981).

144. ETS *Standards*, *supra* note 140, at 10 (1981).

145. *Id.* at 24.

146. *Id.* at 26.

147. ETS *Standards*, *supra* note 140 (1983).

148. Golden Rule Life Ins. Co. v. Mathias, 86 Ill. App. 3d 323, 408 N.E.2d 310 (1980) Transcript of Deposition of Laurel Seneca at 258 (Sept. 23, 1982); *Id.*, Transcript of Deposition of William Kastrinos at 262 (May 5, 1983).

consin[149] showed statistically significant differences in passing rates between White and Black candidates.[150] During the litigation, the *Golden Rule* plaintiffs developed data[151] which at least suggested that in Illinois the disparities in passing rates likewise were substantial.[152] Furthermore, ETS has acknowledged that "[i]t is now widely recognized in the field of testing and measurement that average scores of some groups differ consistently from those of other subgroups on a wide range of standardized tests."[153]

It is plainly impossible to investigate for possible bias in a licensing test without the regular collection and analysis of racial statistics. When, as required by the *Golden Rule* settlement, ETS first released comprehensive data for Illinois test-takers in 1985,[154] the differences in White and Black passing rates were at least as great as the *Golden Rule* plaintiffs had reason to believe. On the life insurance examination, eighty-three percent of the

---

149. Letter dated April 12, 1978 from Stephen F. Heineck, Administrative Officer, Wisconsin Office of Commissioner of Insurance, to Scott Renn, Center for Public Representation, with attachment.

150. With respect to the Wisconsin life examination, the probability of such a distribution of scores happening by chance is less than one and one-half times in 100,00. With respect to the Wisconsin accident and health examination, there was a probability of approximately six in 10,000 of this distribution occurring by chance. Thus, not only were the differences between the passing rates of Black examinees and White examinees statistically significant, they were so extreme as to give rise to an inference of intentional discrimination. A leading text explains: "It has become a convention in social science to accept as statistically significant values that have a probability of occurring by chance 5 percent of the time or less. Many courts have required, or cited with approval, this conventional standard." B. SCHLEI & P. GROSSMAN, *supra* note 70, at 1372 (footnotes omitted). *See* Hazelwood School Dist. v. United States, 433 U.S. 299, 308-09 n.14 (1977)("[a]s a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that teachers were hired without regard to race would be suspect,")(quoting Castaneda v. Partida, 430 U.S. 482, 496-97 n.17.) *But cf.* Watson v. Fort Worth Bank & Trust, 108 S. Ct. 2777 (1988), where a plurality eschewed reference to a "rigid mathematical formula." *Id.* at 2789.

151. Amended Complaint at 25, Golden Rule Life Ins. Co. v. Mathias, 86 Ill. App. 3d 323, 408 N.E.2d 310 (1980) (No. 419-76). Since neither ETS nor the Illinois Insurance Department collected or maintained racial identifying information with respect to applicants for insurance agents' or brokers' licenses, the preliminary study was based on photographs attached to each license application submitted to the Illinois Insurance Department. *See* Affidavit of Suzanne M. Schoolmaster (March 23, 1978) and Affidavit of Donald L. Glick (March 23, 1978) at 2. Golden Rule Life Ins. Co. v. Mathias, 86 Ill. App. 3d 313, 408 N.E.2d 310 (1980). For this reason and others ETS criticized the methodology as unreliable. *See also*, Affidavit of Dr. F. Reid Creeck (April 27, 1978) at 5.

152. The probability of the life insurance examination yielding such differences by chance was approximately one in one hundred (2.65 standard deviations); the probability for the accident and health examination was approximately one in one thousand (3.37 standard deviations). Like the Wisconsin figures, these differences are statistically significant.

153. Brief *Amicus Curiae* of Educational Testing Service at 20, United States v. Texas, No. 85-2579 (5th Cir. Oct. 4, 1985).

154. ETS, Illinois Candidates for Licensure in Life Insurance and Accident & Health Insurance (Apr. 1986).

Whites passed, but only fifty-nine percent of the Blacks.[155] On the accident and health insurance examination, seventy-four percent of the Whites passed, but only forty-one percent of the Blacks. These differences exceed fourteen standard deviations. Since 1.96 standard deviations is statistically significant, and the probability of seven standard deviations is approximately one in one million, these differences are most disturbing.

Not the least important accomplishment of the *Golden Rule* litigation was the discovery that neither ETS nor the responsible Illinois officials had considered it necessary or appropriate to perform the statistical analyses virtually mandated by the *Uniform Guidelines* and the APA *Standards,* not to mention ETS's own internal guidelines. The *Golden Rule* settlement requires that statistics by racial group be compiled, analyzed, and regularly published.

Nonetheless, definitively establishing significant differences in White-Black passing rates simply poses the problem. Some analysts contend that the disparity can be explained in terms of differences in education and in the degree to which candidates are prepared for the tests. Others argue that the examinations are culturally biased against Blacks, a bias having its roots in the fact that most tests are prepared largely by middle-class Whites. Debating these conflicting points of view is not likely to lead to constructive conclusions.

As a matter of public policy, occupational testing should not exclude any qualified persons from engaging in occupations of their choice. If steps can be taken to minimize differences in White-Black success rates without compromising the efficacy of the tests as measures of qualification, those steps should be taken.

V. THE GOLDEN RULE SETTLEMENT

The *Golden Rule* plaintiffs' principal objectives had been to reduce racial differences in performance on the test and to make the test development process open to public scrutiny and accountability. In November 1984, after more than eight years of litigation, the suit was settled. The settlement agreement in large part achieved the objective of public scrutiny and accountability by requiring the annual public reporting of test data and by providing for oversight by an independent advisory committee. The settlement achieved the objective of reducing racial differences in test performance by establishing a procedure for assembling test forms based on selecting test questions which had the least difference in Black-White cor-

155. The data are based on the results of over 9,600 persons who took the life insurance examination, and over 6,600 who took the accident and health examination. *Id.* at Tables I & XI.

rect-answer rates.

The significance of the settlement lies not only in the changes required in the way ETS assembles insurance licensing tests in Illinois, but also in the opportunity provided to measure the extent to which differential racial success rates can be minimized. After examining the principal features of the settlement agreement, the authors assess the legal and psychometric ramifications for occupational licensing brought about by the settlement.

A. The Salient Features of the Settlement

1. Reporting Requirements

The settlement requires that each examinee be asked to provide voluntarily his or her ethnic or racial identification and educational level.[156] This information is to be used for reporting the racial effects of the examination and for guiding the assembly of new test forms.[157] Two public reports are required by May 1 of each year based on data collected during the preceding calendar year. The first, an "examination report," provides "bottom line" information with respect to each part of the examination[158] by race, educational level, and educational level within each racial group.[159] Specifically, the report shows certain data for all examinees, and separately by racial group, educational level, and educational level within each racial group. This information includes the number of examinees, the percentage passing each part of the examination, the percentage passing both parts, and the mean scaled scores and standard deviation on each part. The report also shows the distribution of scaled scores on each part for Black examinees and White examinees with a high school diploma or G.E.D.

The second annual report, an "item report," provides information with respect to each individual test question administered for scoring purposes during the preceding calendar year.[160] Specifically, the item report shows for each question, for all examinees combined and separately for Black ex-

156. Setttlement Agreement and General Release, Golden Rule Life Ins. Co. v. Washburn, no. 419-76 (Cir. Ct. Sangamon County, Ill., settled on Nov. 20, 1984) [hereinafter *Golden Rule Settlement*].

157. *Golden Rule Settlement*, *supra* note 156, at 8-12.

158. The insurance licensing examination is offered in several separate lines of insurance, such as life or accident and health. To be licensed in a particular line, a candidate must pass a two part test. Part 1 is intended to test general product knowledge. Part 2 tests topics specific to each state, such as statutory and regulatory provisions concerning insurance agents. During any given year, several different test forms are administered for parts 1 and 2.

159. *Golden Rule Settlement*, *supra* note 156, at 4-5.

160. *Golden Rule Settlement*, *supra* note 156, at 5-6. The item report includes statistical information on "operational items," i.e., questions used for scoring purposes, but not on "pretest items." *Id.*

aminees and White examinees, the correct answer rates and the r-biserial correlations.[161] The reporting of item statistics permits further analysis to determine whether certain question formats, content areas, or cognitive levels are responsible for a significant portion of any disparity. The item statistics also provide a basis for selecting questions for new test forms so as to diminish, over time, the over-all racial differences.[162]

2. Selection of Test Questions

The most significant provision of the *Golden Rule* settlement establishes a procedure for selecting test questions: when new test forms are assembled, questions in each content area having the least difference in correct answer rates between Black examinees and White examinees must be preferred.[163] The rationale is that, as a matter of fairness, if several questions cover the same content area and are of equal content validity, questions with the least difference in correct answer rates between racial groups should be used. All other things being equal, it is sound public policy to use questions having the least adverse effect on either racial group.

The test assembly procedure required by the settlement has three phases, each of which is carried out *within each content area* as defined by the test's content specifications. Significantly, because the procedure applies *within* each content area, the methodology has no effect on which topics are covered or on the weight attributed to each. Nor does the methodology result in an imbalance of questions available for testing on the various topics. The test developer and user—wholly independent of the settlement— determine the content areas, based on a job analysis or other features of the test development process.[164]

---

161. *Golden Rule Settlement*, *supra* note 156, at 5-6. R-biserials are used to measure the correlation of the performance on an item with performance on a set of items (often the total test score). R-biserial is the correlation coefficient adjusted for attenuation of range.

162. *See infra* text accompanying notes 185-190.

163. *Golden Rule Settlement*, *supra* note 156, at 9-12. The settlement uses absolute values with respect to racial differences. Thus, for example, an item in a given content area with a difference of .05 favoring Whites would be used before an item having a difference of .10 favoring Blacks.

164. A critic of the settlement has argued that "[t]he major problem [with "the *Golden Rule* approach"] involves the distortion of the proportionate weight of subject matter covered by an exam which is established by the content validation process to reflect 'on-the-job' competence requirements." Rebell, *Disparate Impact of Teacher Competency Testing on Minorities: Don't Blame the Test-takers — or the Tests*, 4 YALE L. & POL'Y REV. 375, 394 (1986). Such criticism, however, ignores the fundamental premise of the "*Golden Rule*" methodology: that the classification, ranking, and selection of test questions are carried out *separately within each content area*. *Golden Rule Settlement*, *supra* note 156, at 8-13. Thus, for example, on the 50-item Part 1 of the life insurance examination, ETS uses eighteen different major content areas (many of which are divided into smaller content areas) within which the selection process is *separately* followed. *See* ETS *Multistate Insurance Licensing Program*, Life Test Con-

The first phase, classification, entails sorting questions in each content area into two types (I and II).[165] To be considered Type I, a question must meet two criteria: (1) a minimum correct answer rate for all examinees, White examinees, and Black examinees,[166] and (2) a maximum difference in correct-answer rates between Black examinees and White examinees.[167] Type II questions are questions that fail to meet either or both of these criteria.[168] Following classification of questions into Type I or Type II, the questions are ranked. In each content area, Type I questions are ranked in order by the difference in correct answer rates between Black examinees and White examinees,[169] and Type II questions are similarly ranked.[170] Questions with the least correct answer rate difference are on top, with questions having greater differences placed below.

The third phase of the process, selection of individual questions for assembly into new test forms, involves selecting, within each content area, Type I questions, using first those questions with the least racial difference in correct answer rates.[171] Type II questions are used in a content area *only* when Type I questions are exhausted.[172] Type II questions, when used, are used in order of increasing racial difference in correct answer rates.[173]

With a sufficiently large supply of items in each content area, the process should yield test forms consisting of questions with the least differences in passing rates between racial groups. To assure that the question pool will be ample, the settlement requires that new questions be pre-tested, that is, included in test forms but not counted in the test score.[174] Racial statistics

---

tent Outline (Part I). There can be *no* distortion of content coverage.

165. *Golden Rule Settlement*, *supra* note 156, at 9-10.

166. Specifically, the criterion is that the questions have correct-answer rates for Black examinees, White examinees, and all examinees "not lower than forty percent (40%) at the .05 level of statistical significance . . . ." *Id.* at 10. For a question administered to 1,000 examinees, forty percent at the .05 level is equivalent to a correct answer rate of 36.96 or less. The purpose of this criterion is to assure that the questions which are used have a correct answer rate above the chance or guessing level (25% for a four-option multiple choice question where there is no penalty for guessing).

167. The criterion is that "the correct answer rates of Black examinees and White examinees differ by no more than fifteen (15) percentage points at the .05 level of statistical significance . . . . " *Id.* at 10. For a question administered to 1,000 examinees, a fifteen percentage point difference at the .05 level of statistical significance is equivalent to a difference of approximately 20 percentage points assuming that ten percent of the examinees are Black and the average correct answer rate is .70. The purpose of this criterion is to establish a practical bench mark for distinguishing questions having relatively large racial differences.

168. *Id.* at 10.

169. *Id.* at 10-11.

170. *Id.* at 11.

171. *Id.* at 10-12.

172. *Id.* at 11.

173. *Id.*

174. *Id.* at 13-14. Pretest questions are denominated "Type III Items" in the *Settle-*

are collected on the pre-test items[175] and, when test forms are assembled, pre-test questions are included in the pool of questions which are classified as Type I or II, ranked within the appropriate content area in accordance with the racial differences in correct answer rates and assembled into test forms to the extent they rank higher than questions previously used as operational items.[176] With the continued pre-testing of questions, questions assembled into test forms will have progressively fewer differences in correct answer rates between racial groups, at least until a point of irreducible difference is achieved. Indeed, one of the more interesting empirical questions which the settlement may illuminate is the extent to which it is possible to eliminate or reduce such racial differences.

By using statistical criteria, the settlement recognizes that, even after questions which have the greatest disparities in correct-answer rates between racial groups are identified, it may not be possible to explain such disparities on the basis of a facial analysis of the questions. There may well be questions having relatively large differences in the correct answer rates between racial groups, yet experts from various fields—multi-cultural linguists, insurance practitioners, and test developers--cannot identify any element in content, format, or language which would account for the differences.

3. Other Provisions

Under the settlement, every other year a test form and answer key for each part of each examination must be made public. As with the Truth-in-Testing legislation in New York,[177] the objective is to provide public accountability and to help prospective examinees prepare for the examination.

The settlement sets additional standards. A reading level standard requires that the reading level of the examinations be at the high school level.[178] In addition the examinations must adhere to the APA *Standards*.[179] By adopting the APA *Standards,* the settlement requires, at a minimum, that the examinations be content-valid[180] and that they be shown to be job-related based upon a job analysis.[181] The requirement of content

---

ment. Id. at 13.

175. *Id.* at 14.

176. *Id.* at 14-15.

177. New York Educational Testing Act of 1979, N.Y. EDUC. LAW §§ 340-348 (McKinney 1988).

178. *Golden Rule Settlement*, *supra* note 156, at 7-8.

179. *Id.* at 6-7. The *Settlement* also requires that ETS adhere to its "Standards for Quality & Fairness." *Id.* Even prior to entering into the settlement, ETS had claimed to adhere to the APA *Standards*, *see supra*, text accompanying note 140.

180. APA *Standards*, *supra* note No. 11.1.

181. APA *Standards*, *supra* note 17, Nos. 1.6, 11.1 & Comment to Standard No. 11.1

validity exists even in the absence of a showing of adverse racial impact. As indicated above,[182] the APA *Standards* require differential validation studies (a showing that the examinations are valid for different racial groups) where prior research indicates that racial groups perform differently.[183]

B. Legal Significance of the Settlement

The *Golden Rule* settlement establishes detailed standards, procedures, time-tables and a system for monitoring compliance. By entering into the settlement, ETS seemingly recognized that it is feasible to minimize racial differences in correct answer rates. The settlement establishes a standard against which acts and omissions of test developers and users may be judged. The settlement may be viewed as a "less-discriminatory alternative" to other testing methods, thereby allowing plaintiffs to establish liability even where defendants have shown that a test is content valid. Further, to the extent the settlement procedures have become known in the testing industry,[184] the knowing failure to use them might constitute evidence of intentional discrimination.

In addition to its potential use as a "sword," the settlement may serve as a "shield." Employers, government agencies, and other entities subject to anti-discrimination laws (Title VII, constitutional standards, or state civil rights laws) may well reduce their exposure to suit by taking appropriate steps to reduce adverse racial impact. Using a procedure for selecting test questions having the smallest racial differences in correct answer rates should decrease such impact.

C. Psychometric Significance of the Settlement

1. Effect of the Settlement in Reducing Racial Differences

The annual reports prepared by ETS pursuant to the *Golden Rule* set-

---

at 64. *Cf.* APA *Standards*, *supra* note 17, No. 10.4.

182. *See supra*, note 141 and accompanying text.

183. APA *Standards*, *supra* note 17, No. 3.10. *See also* Comment to Standard No. 3.10 at 27 and Introduction at 2-3.

184. The settlement has been discussed in an official publication of the American Psychological Association. Anrig, *'Golden Rule': Second Thoughts*, 18 APA MONITOR No. 1 at 3 (Jan. 1987). It has also been the subject of questioning by the APA's Committee on Psychological Tests & Assessment. Committee on Psychological Tests & Assessment, American Psychological Assoc., *Implications for Test Fairness of the 'Golden Rule' Company Settlement* (August 11, 1988). ETS and Golden Rule also have debated the settlement in the literature. *See* Rooney, *Golden Rule on Golden Rule*, 6 EDUC. MEASUREMENT: ISSUES & PRAC. No. 2 at 9 (Summer 1987); Anrig, *ETS on "Golden Rule,"* 6 EDUC. MEASUREMENT: ISSUES & PRAC., No. 3 at 24 (Fall 1987); Rooney, *A Response From Golden Rule to "ETS on 'Golden Rule',"* 6 EDUC. MEASUREMENT: ISSUES & PRAC. No. 4 at 19 (Winter 1987).

tlement contain data permitting a comparison of the performance of Black and White candidates on traditionally-assembled (i.e., pre-settlement) test forms and on forms assembled under the settlement. The settlement specifies that, within each content area, items are to be selected in order of increasing differences in Black-White correct answer rates. In practice, however, ETS has apparently given lower priority to this requirement than to its traditional test-assembly criteria. The selection of test items with the least Black-White differences within each content area has therefore been affected by other considerations: (1) providing specific numbers of items within each of three "cognitive levels"—knowledge, application/analysis, and evaluation; (2) providing specific numbers of items within each of four "item types"—negative stems, classification sets, roman numeral format, and four-choice positive stems; (3) satisfying a predetermined distribution of item difficulty levels; and (4) satisfying minimum value criteria for proportion-correct and r-biserial statistics.

In addition to these factors, which tend to mitigate the effects of the settlement, the actual item pool is far smaller than originally anticipated because of asserted security breaches, ETS's decision to release a test form containing previously viable items, and changes in the law which rendered obsolete many state items. These actions, as well as others,[185] served to work against the reduction in Black-White differences. However, despite the relatively low priority apparently accorded by ETS to reducing racial differences, the differences in Black-White correct answer rates and in passing rates have been statistically significantly reduced.

Test forms assembled according to the terms of the settlement were first introduced in 1986, but were not used exclusively as Part 1 forms until 1987. The passing rates of Black and White examinees on the Part 1 tests for 1985 and 1987[186] provide a basis for determining the effects of the settlement upon the disparity between Black and White passing rates. The annual reports contain data showing the number of examinees and the passing percentages rounded to the nearest one percent, but not the precise number of examinees passing. Data are reported for examinees from three educational levels, each of which contained a minimum of 100 examinees from each racial group. The pertinent data are as follows:

---

185. For example, some forms were "double-equated," so that approximately 70% of the items selected had appeared in previous forms. The criteria for selecting equating items were given higher priority than the settlement's requirement of choosing items with the least racial difference in correct-answer rates.

186. Data for 1986 were not used in this analysis because during that year both pre-settlement (traditional) forms and post-settlement ("Golden Rule") forms were administered.

TABLE 1

	1985 Traditional Forms		1987 Golden Rule Forms	
	Number of Examinees	Percentage Passing	Number of Examinees	Percentage Passing
LIFE INSURANCE EXAM				
White				
HS/GED	1608	81	1635	84
Some Coll.	2879	88	3496	90
4yr Deg.+	2256	95	3646	96
Black				
HS/GED	194	56	236	57
Some Coll.	537	65	617	75
4yr Deg.+	283	82	455	86
ACCIDENT & HEALTH INSURANCE EXAM				
White				
HS/GED	1011	73	1491	73
Some Coll.	2021	82	3045	81
4yr Deg.+	1794	91	3291	90
Black				
HS/GED	140	36	199	36
Some Coll.	375	48	533	57
4yr Deg.+	188	67	423	79

These data can be statistically analyzed by converting the passing rates into the number of examinees passing and failing the examination within each group. The most conservative method of analysis was used: (1) the number passing in 1985 was calculated as the largest number which would satisfy the passing percentage; (2) the number passing in 1987 was calculated as the smallest number which would satisfy the passing percentage; (3) the Mantel-Haenszel equation was used for each educational group separately within each Part 1 test;[187] and (4) a correction for continuity was

187. The Mantel-Haenszel Chi-squared provides for the calculation, within each subgroup of data, of the variance and of the difference between observed and expected values. The Chi-squared statistic is computed as the square of the sum, over all subgroups of data, of the differences between observed and expected values divided by the sum, over all the subgroups of data, of the variances. N. Mantel & W. Haenszel, *Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease*, 22 JOURNAL OF THE NATIONAL CANCER INSTITUTE 719 (1959).

subtracted.[188]

Using this conservative approach, which also controlled for educational level, Black passing rates improved statistically significantly by 4.532 standard deviations whereas White passing rates decreased insignificantly by 0.697 standard deviations.[189] Thus, there has been a significant decrease in the gap between Black and White passing rates since the adoption of the test assembly procedures required by the settlement.[190]

2. The Effect of the Settlement on the Psychometric Quality of the Examinations

To determine the effect the settlement is having on the psychometric quality of the examinations, regression analyses were performed.

188. The correction is applied by subtracting 0.5 from the absolute value of the sum, over all the subgroups of data, of the differences between observed and expected values.

189. The number of standard deviations is equal to the square root of the Mantel-Haenszel Chi-squared.

190. Further, comparing traditional forms with "Golden Rule" forms, there has been a statistically significant decrease in the average difference between Black and White correct answer rates.

Table 2

Illinois Life Insurance Licensing Examination
UL1-UL6 Pre-Golden Rule
UL7-UL8 Golden Rule - Social Security Item Pool
UL9-UL10 Golden Rule - 1985 Item Pool
SL1-SL3 Pre-Golden Rule
SL4 Golden Rule

1986 Form	W p^+ mean	W p^+ est SD	W rbis mean	W rbis est SD	B p^+ mean	B p^+ est SD	B rbis mean	B rbis est SD	T rbis mean	T rbis est SD	W-B p^+ mean	W-B p^+ est SD
UL1	.691	.174	.450	.130	.590	.190	.381	.154	.467	.127	.101	.070
2	.680	.193	.413	.139	.595	.201	.400	.165	.437	.151	.085	.053
3	.701	.161	.432	.123	.588	.167	.424	.138	.452	.126	.113	.068
4	.701	.172	.429	.143	.608	.186	.413	.168	.450	.148	.093	.069
5	.709	.169	.475	.138	.575	.180	.441	.157	.491	.122	.133	.057
6	.736	.143	.485	.141	.620	.170	.400	.151	.490	.142	.117	.090
7	.720	.021	.439	.133	.620	.167	.387	.151	.451	.133	.100	.065
8	.725	.158	.434	.128	.625	.176	.426	.152	.458	.130	.099	.063
9	.697	.180	.446	.144	___	___	___	___	.455	.136	___	___
10	.707	.186	.430	.111	.639	.198	.402	.183	.437	.115	.068	.063
SL1	.675	.176	.409	.147	.584	.172	.391	.143	.419	.147	.091	.065
2	.636	.197	.425	.118	.537	.192	.411	.124	.437	.116	.100	.062
3	.627	.205	.426	.149	.539	.193	.376	.148	.423	.142	.087	.072
4	.761	.140	.448	.126	.686	.158	.485	.142	.479	.128	.075	.049

	p^+			rbis				
	r	a	b	r	a	b	Nw	Nb	Nt
UL1	.930	.190	.849	.706	.223	.595	1016	127	1350
2	.965	.130	.923	.734	.165	.621	1314	193	1789
3	.915	.180	.886	.681	.175	.605	1393	220	1865
4	.929	.180	.858	.705	.182	.599	966	158	1333
5	.948	.196	.892	.574	.251	.507	821	118	1670
6	.847	.296	.710	.673	.234	.627	955	133	1288
7	.921	.202	.835	.800	.167	.703	1173	178	1565
8	.934	.201	.837	.730	.171	.617	1321	209	1785
9	___	___	___	___	___	___	644	95	872
10	.948	.136	.895	.514	.305	.311	687	105	947
SL1	.930	.120	.949	.810	.085	.828	2425	357	3876
2	.950	.111	.979	.822	.104	.782	2661	412	3595
3	.936	.090	.995	.717	.155	.721	1223	189	1666
4	.953	.185	.840	.517	.225	.460	1197	173	1588

Figures 3 and 4 contain graphic representations of the regressions for two Part 1 Life forms administered in 1986,[191] a traditional form (Form 1) and a *"Golden Rule"* form (Form 10), for which summary statistics are presented in Table 2.[192] The top portion of Figure 3 contains the scatterplot for Form 1 of the White proportion correct (ordinate) and the Black proportion correct (abscissa). Each individual item is represented by a point on the scatterplot. The two lines drawn on each scatterplot depict (1) the regression line which best fits the points and (2) the diagonal line on which all points would lie if White and Black item performances were equal. The bottom portion of Figure 3 contains the comparable scatterplot for Form 10. Figure 4 contains the scatterplots of the White and Black r-biserial correlation coefficients for the items on Form 1 (top) and Form 10 (bottom).

191. Data from 1986 are used because during that year both traditional and "Golden Rule" forms were administered, thereby making comparisons readily available.

192. The Table contains the following data for each form: (1) the mean proportion of correct answers (p^+) on each item for White candidates; (2) the standard deviation of the proportion of correct answers on the items of the test for White candidates; (3) the mean and standard deviation of the r-biserial correlation coefficients of the test items for White candidates; (4) the same statistics for Black candidates; (5) the mean and standard deviation of the r-biserial correlation coefficients for the test items for the total group of candidates; (6) the mean and standard deviation of the difference between White and Black correct-answer proportions on the items of the test; (7) the correlation coefficient *r*, intercept (a), and slope (b) of the White and Black correct answer proportions on the items of the test; (8) the correlation coefficient, intercept, and slope of the White and Black r-biserial correlation coefficients on the items of the test; and (9) the number of White, Black, and total test takers on each test form.

Figure 3

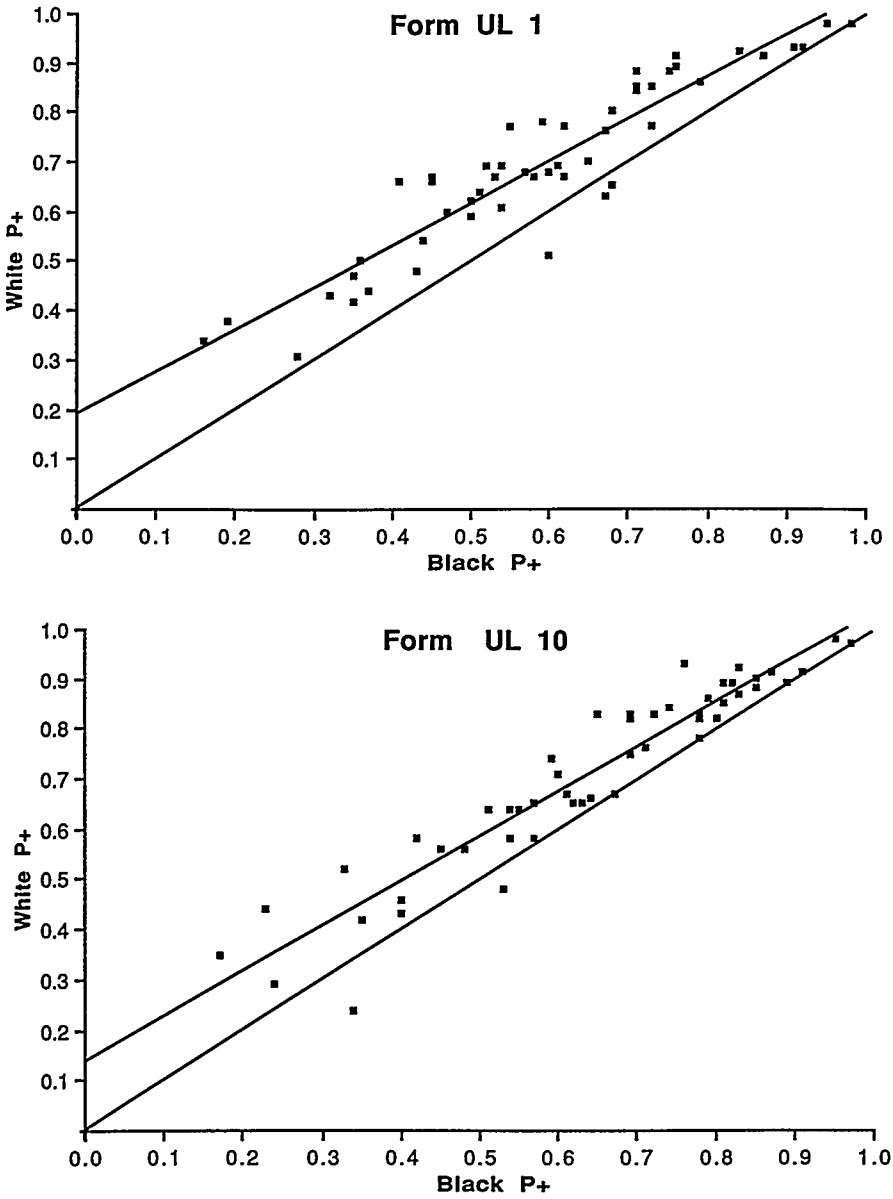
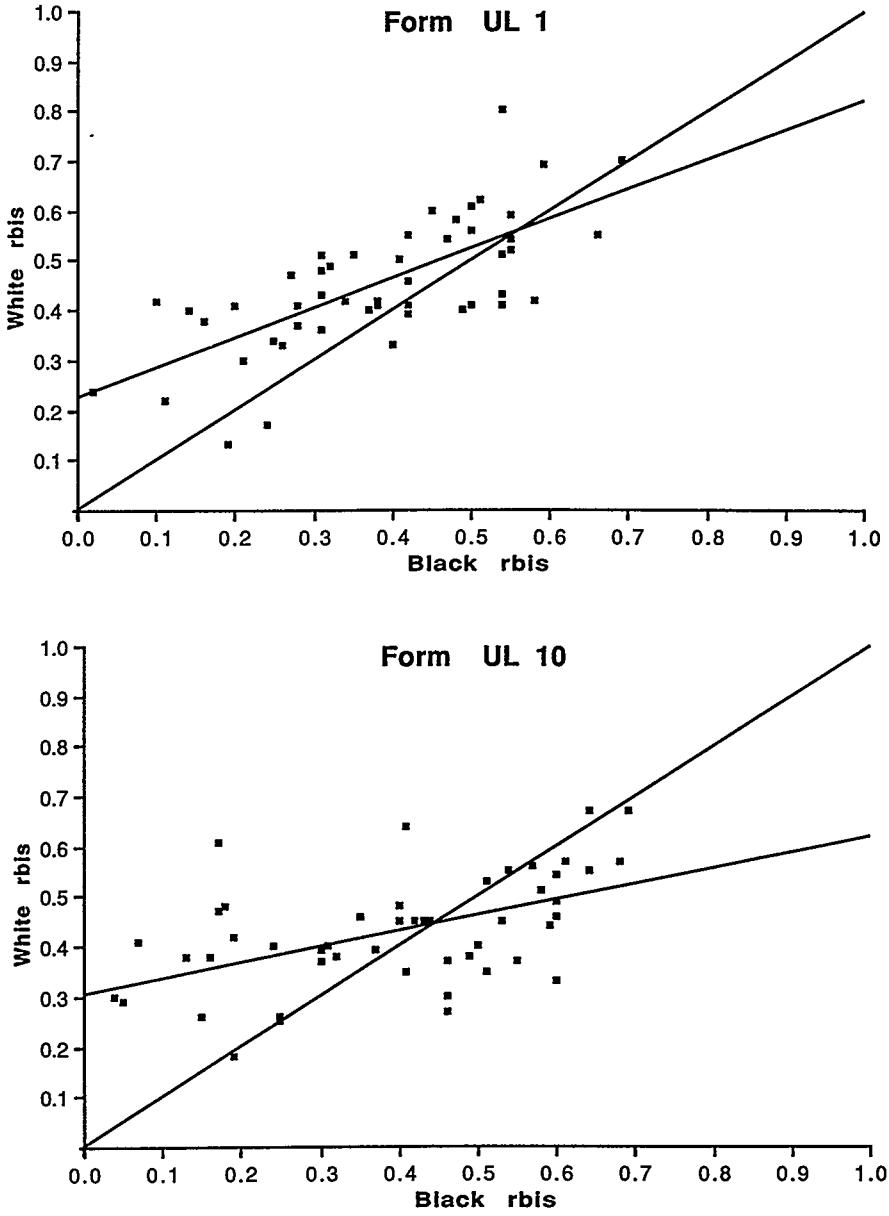

Form UL 1
White P+
Black P+
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Form UL 10
White P+
Black P+

Figure 4


Form UL 1
White rbis
Black rbis
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Form UL 10
White rbis
Black rbis

A comparison of the top and bottom portions of Figure 3 reveals that the relationship between White and Black proportions correct is very orderly on both forms, as a reflection of the large correlation coefficients, with the points lying relatively close to the regression lines. The two graphs do differ in the relationship between the diagonal line and the regression line, as a reflection of the decreased White-Black difference on Form 10, with the points for Form 10 lying generally closer to the diagonal line and the regression line lying much closer to the diagonal line.

A comparison of the top and bottom portions of Figure 4 reveals that the correlation between White and Black r-biserial correlation coefficients is somewhat reduced for Form 10 as compared to Form 1; however, as indicated by the points in the scatterplots, the actual values of the r-biserial correlation coefficients are not different for either White or Black candidates between Form 1 and Form 10. In addition, an examination of Table 2 reveals, with respect to "*Golden Rule*" and "pre-*Golden Rule*" forms, that the means and standard deviations of the r-biserial correlation coefficients are unaltered by the introduction of the "*Golden Rule*" forms, for White candidates, for Black candidates, or for the total set of candidates. It is clear from the data that the implementation of the settlement has not adversely affected the psychometric properties of the examinations. ETS shares this view.[193]

---

193. Anrig, *ETS on "Golden Rule,"* 6 EDUC. MEASUREMENT: ISSUES & PRAC. No. 3, 24 (Fall 1987).

This is not to say that there have not been misdirected psychometric attacks on the settlement. For example, Linn and Drasgow, *Implications of Judicial Decisions for Test Construction*, 6 EDUC. MEASUREMENT: ISSUES & PRAC. No. 2, 13 (Summer 1987), base their criticism on an *assumed* non-demonstrable difference in underlying ability between Blacks and Whites. *Id.* at 15. *See also supra* note 164.

ETS researchers ran two experimental sections of the Graduate Record Examination (GRE) in which they selected items within each content area which minimized Black-White differences. Although reaching somewhat negative conclusions as to the desirability of using "impact reduction techniques"— unquestionably prompted by the *Golden Rule Settlement*—the researchers recognized that:

> First, such techniques *can* reduce impact. . . Second, the resulting tests *can* be made to look parallel in form and content to conventionally constructed tests and meet their content specifications if the item pools are sufficiently large. Third, the *average* difficulty level of the resulting tests *can* be maintained without changing current test development procedures for adhering to average difficulty specifications. However, the *distribution* of item difficulties will change. . . This may be a controllable phenomenon. . . .

Hackett, *Test Construction Manipulating Score Differences Between Black and White Examinees: Properties of the Resulting Tests* at 31 (ETS, July 1987)(emphasis in original).

The ETS researchers note that one method of control involves requiring a specific distribution of item difficulties (a requirement to which ETS adhered in constructing the Illinois insurance examinations). Another method of control, not noted by the researchers, would involve reducing the use of very high difficulty items (a requirement which was incorporated in the *Golden Rule Settlement*).

3. Expanding Use of the *Golden Rule* Methodology

The test-assembly principles of the settlement have been accepted in their essentials by a national association of teacher-college educators.[194] Legislation has been introduced in several states to extend these principles.[195] In essence, the bills would require the collection, analysis and public reporting of racial data and the selection of questions for test forms having the least racial difference in correct-answer rates.[196] In New York, legislation has been enacted which establishes a mechanism for collecting and analyzing data on differential racial performance.[197]

ETS has resisted any extension of the *Golden Rule* settlement principles.[198] Indeed, ETS's president has confessed that he "made a mistake" in approving the settlement,[199] in part, at least, because ETS did not foresee that Golden Rule and others would seek to extend the methodology to other types of examinations and to other states. ETS appears to believe it necessary, in opposing any extension, to characterize the *Golden Rule* settlement as an *ad hoc* solution to a unique problem.

ETS further considers that the *Golden Rule* procedure is based on a faulty premise—"that group differences in performance on test questions primarily are caused by 'bias.' "[200] No such premise underlies the settlement, since it makes no assumptions about "bias." Rather, its premise is that if there are racial differences at the item level which can be reduced by using questions having smaller racial differences *in the same content areas which test the same material at similar levels of cognition,* fundamental fairness requires that such differences be reduced, regardless of whether the differences are caused by "bias" or by other factors which are not easily measured. In effect, the settlement provides a means of reducing *unnecessary* racial differences.

While expressing displeasure with the settlement, ETS "supports [the] use of appropriate statistical analyses of group differences in the process of

194. The settlement has been supported in principle by the American Association of Colleges for Teacher Education. Daily, *The Appropriate Role of Testing in the Teaching Profession?* in WHAT IS THE APPROPRIATE ROLE OF TESTING IN THE TEACHING PROFESSION? 49, 55 (1987).

195. Wis. Assembly Bill 855 (1985-86 Legis.); Mass. Sen. Bill 758 (Jan. 1986); Calif. Assembly Bills 4045-46 (1985-86 Reg. Sess.).

196. *See supra* note 195.

197. N.Y. Educ. Law §§ 341-a-346-a (McKinney 1988). This Legislation was approved on June 29, 1987.

198. *See, e.g.*, Statement by ETS With Respect to Wis. Assembly Bill 855, before Wis. Assembly Committee on Financial Institutions & Insurance (Mar. 3, 1986).

199. Anrig, *'Golden Rule': Second Thoughts*, 18 APA MONITOR No. 1 at 3 (Jan. 1987).

200. *Id.*

selecting questions for tests."[201] It has identified one of these methods as "differential item functioning."[202] ETS began using this method in 1987 in insurance licensing examinations (in states other than Illinois) and in teacher tests in the National Teachers Examination (NTE) programs.[203] ETS's position appears to be that differential item functioning is a preferable means of identifying "biased" questions.[204] ETS proposes to compare item responses of persons who received the same total score on the examination in order to determine if one or more individual questions might be biased.[205]

Notwithstanding ETS's sponsorship, the differential item functioning methodology is not an acceptable alternative to the *Golden Rule* approach. The method of matching examinees on their total test score presumes that (1) the total test score is a valid measure of knowledge of the subject matter tested, and (2) there is not even one biased question on the test which would have resulted in examinees against whom the item is biased receiving lower scores. The methodology by which ETS proposes to "confirm the fairness of test questions" must therefore presume the collective fairness of all test questions.

As has been noted, item analysis techniques can detect biased items only if the degree of bias is greater for one or more items than the bias of the other items on the examination. In discussing item response theory us-

---

201. *Id.*
202. *Id.*
203. *Id.*
204. *Id.*; Letter dated November 21, 1985 from George Elford, Director, ETS Midwestern Regional Office, to Hildegard Neujahr, Director of Administrative Services, Wisconsin Insurance Commission [hereinafter "Elford letter"].
205. As ETS has explained:

> The procedure begins with a very straightforward premise: If a test question is fair, people who know the same amount about the subject being tested should have an equal chance of answering the question correctly, regardless of race, ethnicity, gender, background and the like. For example, women and men with equal knowledge of mathematics would be expected to do equally well on a fair mathematics question.
>
> The first step in this procedure is therefore to identify clusters of people in the various subgroups to be compared, who are matched in their knowledge of the subject being tested. Estimates based on standardized test scores have proved to be an available, accurate and reliable means for matching knowledge and ability. Thus, test-takers from different subgroups can be grouped in clusters based on their scores on the test as a whole.
>
> The next step is to calculate how hard each question is for the members of each subgroup within each cluster to be compared. Do women and men with comparable scores on the test as whole do equally well on each question? The procedure yields a numerical index for each question reflecting any differences in difficulty between subgroups within each cluster. The index augments well-established test fairness review procedures by flagging questions for intensive additional review. Based on this review, any questions which may be unfair can be revised or eliminated.

Statement enclosed with Elford letter, *supra* note 204.

ing what are called delta plots, Frederic Lord of ETS conceded that "we could say that the items are all equally biased or, conceivably, equally unbiased."[206] The differential item functioning method does nothing to eliminate this problem; it presumes that the total test score is equal to knowledge, and it makes no allowances for the possibility of biased items. For example, if each item were ten percent biased (i.e., one group with equal knowledge scored ten percent lower than another group), then the examinees against whom the items were biased would be compared to examinees in the other group who had ten percent less knowledge, but the method would identify no biased items because it would reflect equal performances by both groups.

ETS's methodology matches examinees who receive the same total test scores. However, if at least one item is biased, one group of examinees cannot obtain the same total test score as another group even if both groups have the same knowledge of the subject matter, unless other items are compensatingly biased in their favor. Such a compensating situation where biased items offset each other is unlikely.

The statistical assumptions underlying differential item functioning are satisfied only when the controlled factor, such as age in a smoking study, is independent of the variable tested, such as the effect of smoking on death rates.[207] Applied to the test bias problem, it is clear that one cannot rationally assume that total test scores are independent of individual item performance; the test score is nothing but the sum of the item scores. Further, ETS does not specify the level of statistical significance used in connection with the differential item functioning methodology for identifying biased questions. It is therefore impossible to predict—assuming the other flaws did not exist—whether the methodology would be of any practical use in identifying biased questions.

More fundamentally, however, ETS's approach uses the methodology merely to *identify* biased items which are then subjected to some unspecified "intensive additional review." The approach thus relies on an apparently standardless, subjective means of ultimately determining whether a question is biased. The method assumes that it is possible to make an ultimate determination of whether a question is biased by a facial review of the question.

The *Golden Rule* settlement does not rely on a subjective review of questions, since in many, if not most, cases, bias cannot be detected in that

206. F. LORD, APPLICATIONS OF ITEM RESPONSE THEORY TO PRACTICAL TESTING PROBLEMS (ETS 1980) at 213.

207. *See, e.g.*, E. Frome & H. Checkoway, *Use of Poisson Regression Models in Estimating Incidence Rates and Ratios*, 121 AM. J. OF EPIDEMIOLOGY 309 (1985)(Mantel-Haenszel is "strictly suitable only when there is no interaction between relative risk and the covariate . . . . ").

fashion. Statistical measures afford the sole objective and reliable means not only of identifying possibly biased questions, but also of determining whether a particular question should be used. If a statistical measure suggests bias, a question should not be used when other questions having less adverse impact on a particular group are available. The *Golden Rule* settlement criteria are thus objective and self-executing.

VI. CONCLUSION

The testing industry's current practice of using only content validation procedures and its wholly relative method of detecting bias preclude proving the existence or absence of test bias. Likewise, current legal standards for determining liability for test bias involve allocating burdens of proof which may be result-determinative.

As a practical matter, a plaintiff may be able to satisfy the Title VII adverse impact standard where he or she would be unable to satisfy the constitutional standard. Even so, Title VII plaintiffs must persuade the courts that the statute applies to licensing and reaches private test developers.

Quite apart from legal proceedings, the technical obstacles to proving or disproving bias may be avoided by directly addressing the practical question: can the purported purpose of a test be met if test developers take steps to minimize racial differences to the greatest extent possible? Such an approach avoids the admitted inability of psychometricians to detect bias in the absence of a valid, independent external criterion.

While the legal and psychometric issues relating to racial bias in occupational testing continue to be in dispute, the *Golden Rule* settlement avoids much of the debate. The settlement is a practical way of mitigating racial differences without adversely affecting test content or validity. It makes possible the development of both fairer test, and tests less subject to legal challenges.

The *Golden Rule* settlement is certainly not the last word, but it may help focus

> the debate away from a strictly academic discussion of the issue of racial bias toward a debate about practical solutions. If the settlement goes beyond that, and succeeds in pushing ETS, other test developers, and state regulators and legislators in the direction of establishing meaningful and accountable means of reducing unnecessary racial differences, it will have made a major contribution toward ensuring equal opportunity for all.[208]

208. Rooney, *A Response From Golden Rule to "ETS on 'Golden Rule',"* 6 EDUC. MEASUREMENT: ISSUES & PRAC. No. 4 at 19, 23 (Winter 1987).