Berkeley La Raza Law Journal

Volume 12 *Number 2 (2001)* Article 10

2001

A Psychometric Model for Preserving Discrimination

Martin Shapiro

A Psychometric Model For Preserving Discrimination†

Martin Shapiro‡

The business of developing admission tests may be divided into four parts, each of which will be discussed in turn. The first step is defining the content domain of the test. The second step is writing and selecting test items which represent the specified content domain and which satisfy technical standards for test reliability. The third step is deriving a mathematical equation which describes the relationship between test scores and criterion measures. The fourth step is determining the predictive validity of the mathematical equation by testing the equation to a new group of applicants. The process is never-ending. The integrity of the admissions process necessitates secrecy, limiting test items and test forms to finite [periods of use]. The deletion and addition of test items and the construction of new test forms require reiterations of the process. Additionally, each new test form must be equated to previous forms.

I.
CONTENT DOMAIN

Admissions tests have no predetermined content. Unlike achievement tests the content of an admissions test is not prescribed by either course content or subject matter content. The test may be designed to measure any knowledge, or skill or ability which the test manufacturer presumes is predictive of some selected measure of future performance.

Content domain is the "set of behaviors, knowledge, skills, abilities, attitudes or other characteristics to be measured by a test, represented in a detailed specification, and often organized into categories by which items are classified."[1] Defining the content domain of the test is a subjective judgmental process. Dissatisfaction with a relatively insufficient mathematical relationship between test scores and criterion may, and should, prompt attempts to modify the content domain of the test, but the choice of modifications will be a product of subjective judgment. Or, the content domain of the test may be modified, in spite of a relatively sufficient relationship between test scores and criterion measures, because the test has consequences which are deemed undesirable. For example, the empirical observation that men attained higher PSAT [Preliminary Scholastic Aptitude Test] scores than women was deemed undesirable because awards of National Merit

---

† Expert report submitted on behalf of Intervening Defendants (Student Intervenors), Grutter v. Bollinger, 137 F. Supp. 2d 821 (E.D. Mich. March 27, 2001) (No. 97-75928).

‡ Martin Shapiro is Professor of Psychology at Emory University.

1. Joint Committee on Standards for Educational and Psychological Testing, *Standards for Education and Psychological Testing* 174 (1999).

Scholarship [are] based, at least in part, on PSAT scores. To reduce this observed difference between male and female scores, a writing section was added to the PSAT under the presumption that women would attain higher writing scores than men, thereby reducing the gap between their scores. The choice of content domain for a test may be driven in part by the consequences of the test. Societal, political values and judgments affect the content domain of a test.

There are no statistical criteria for measuring the adequacy of the content domain of a test. If test scores have a relatively sufficient relationship to criterion measures, there is little or no incentive to investigate alternative content domains. Particularly in situations where there are no competing admissions tests, such as the situation which currently exists with respect to the LSAT [Law School Admissions Test], MCAT [Medical College Admissions Test] and GRE [Graduate Records Examination], there are the disincentives of cost to the manufacturer and unfamiliarity to the consumer which operate to discourage revising the content domain. If the perception of the educational institutions is that the test is fulfilling its purpose, maintaining the status quo is the best strategy. Barring intervention by a powerful third party, e.g., the United States government in the above example of the PSAT, the content domain of an admissions test which is perceived to be fulfilling its purpose is highly resistant to change. A change in content domain creates an awkward transition period during which applicants tested with old test forms and applicants tested with new test forms are competing in the same admissions cycle.

II.
ITEM SELECTION

A test is the sum of its items. The impact of a test upon any group of individuals is the sum of the impact of the individual test items. As explained below, the typical method of selecting test items is designed to systematically maintain and enhance test-score differences between groups of individuals.

The test manufacturer must continuously replenish the inventory of untested items for each of the categories of items on the test. From this untested item pool, items are first subjectively chosen for inclusion on test forms as pre-test items. Candidate performance on the pre-test items is not counted in a candidate's test score. The pre-test items are evaluated for their suitability as future test items and the performance of the candidates on each pre-test item provides the statistical information about pre-test items upon which subsequent item selection decisions are made. The pre-test items are evaluated in terms of item difficulty and item reliability. The subsequent objective selection of pre-test items as operational items on subsequent test forms must satisfy the content domain of the test and must meet the distribution of item difficulty dictated by the requirements of test-form equating. Once the content and difficulty criteria are satisfied, pre-test items are selected in descending order of their reliability. Of course in practice, there is a considerable amount of discretion involved in the item selection process. For example, if there are no sufficiently reliable items with a targeted difficulty within one content category, the test assembler may select a more difficult item within that content category and offset the change in difficulty with the selection of a less difficult item within another content category. Also, the test specifications may allow some leeway in the number of test items required within each category of test items, a situation in which greater discretion is permitted in selecting items.

Item difficulty in its rawest form is simply the proportion of candidates who give the right answer. In practice, item difficulty is often transformed into a standardized measure which is designed to correct for the ability of the particular group of candidates taking the test at a particular time. The proportion of candidates who give the right answer on any one test item may be compared to the proportion of candidates who give the right answer on other test items. The comparison may be conducted across all candidates or the comparison may be conducted within a subset of candidates who received the same total test score on the other items or who received the same aggregate test score on a specific subgroup of the other items. If the difficulty level of the other items is available from previous test forms on which the other items occurred either as pre-test items or as operational items, the comparison may be conducted within a test form. If the difficulty level of the other items is not available from previous test forms, the comparison may be conducted across test forms which share common items, either as pre-test items or as operational items.

For purposes of item selection, reliability refers to the internal consistency of candidate performance on the individual test items. In its rawest form, if a right answer is scored as a 1 and a wrong answer is scored as a 0, the reliability of an item is simply the correlation between the score on that one item (0 or 1) and the sum of the scores on the other N-1 items, where N is the total number of items on the test. The sum of the scores on the other N-1 items may be calculated as the sum of scores on all N items, i.e., the total test score, minus the score on the one item whose score is being correlated with the total test score. Or, the sum of the scores on the other items may be calculated as the sum of scores on the other items within a subgroup of test items, i.e., a sub-test score excluding the one item whose score is being correlated with the sub-test score. An "internal consistency coefficient" is an "index of the reliability of test scores derived from the statistical interrelationships of responses among item responses or scores on separate parts of a test."[2] The same concept of internal consistency is incorporated into "item response theory (IRT)."[3] In the *Standards*, item response theory is defined in rather technical language as a

> "mathematical model of the relationship between performance on a test item and the test taker's level of performance on a scale of the ability, trait, or proficiency being measured, usually denoted as θ [theta]. In the case of items scored 0/1 (incorrect/correct response) the model describes the relationship between θ and the item mean score (P) for test takers at level θ, over the range of permissible values θ. In most applications, the mathematical function relating P to θ is assumed to be a logistic function that closely resembles the cumulative normal distribution."[4]

In other words a good test item is one which less-able candidates get wrong and more-able candidates get right. Good items come in a range of difficulties. Good

2. *Id.* at 176.

3. *Id.* at 177.

4. *Id.*

items which are less difficult require less ability and good items which are more difficult require more ability. Lacking an independent, external measure of ability, a candidate's ability is defined as the candidate's test score. The logic has come full circle. A good test item is an item that is consistent with the other items. A good test item is an item which high scorers get right and low scorers get wrong.

The item selection criteria are designed to select the best pre-test items and to discard the worst pre-test items, within the limitations imposed by the required content of the test and the targeted difficulty level. The item selection criteria are designed to maximize the ability of the admissions test to distinguish, i.e., to discriminate, among applicants of different ability as defined by their admissions test score. There is no absolute standard for admission. There is a desired or practical class size and the institution evaluates applicants relative to each other within the limits imposed by that class size.

> "The test developer is also responsible for ensuring that the scoring procedures are consistent with the purpose(s) of the test and facilitate meaningful interpretation. The nature of the intended score interpretations will determine the importance of psychometric characteristics of items in the test construction process. For example, indices of item difficulty and discrimination, and inter-item correlations, may be particularly important when relative score interpretations are intended. In the case of relative score interpretations, good discriminations among the test takers at all points along the construct continuum is desirable. It is important, however, that the test specifications are not compromised when optimizing the distribution of these indices."[5]

In seeking items that maximally discriminate between ability levels, i.e., maximize the range of raw test score differences between the highest and lowest scoring candidates, the test manufacturer continues an irreversible process. To explain this process, consider the first group of test candidates to be identified by some shared distinguishing personal characteristic(s) and a shared propensity to obtain higher scores on the admissions test and consider the second group of test candidates to be identified by some other shared distinguishing personal characteristic(s) and a shared propensity to obtain lower scores on the admissions test. Pre-test items which the higher-scoring group answers correctly and the lower-scoring group answers incorrectly are considered to be good items and are likely to be included as operational items on a subsequent test form. The same process is repeated on each subsequent test form. With a sufficiently large pool of pre-test items, the resulting difference between the scores of the higher-scoring group and the lower-scoring group is limited only by the difference between the average correct answer rate of the higher-scoring group and the correct guessing rate of the lower-scoring group. With or without remedial learning for the lower-scoring group, the difference between the test scores of the higher-scoring and the test scores of the lower-scoring group approximates a negatively accelerated, positive monotonic function, i.e., a never decreasing process which approaches some upper limit (asymptote) in increasingly smaller steps. Remedial learning for the lower-scoring

5. *Id.* at 39-40.

group would not reverse the process, because the selection process for pre-tested items continues to discard items which would reduce the test-score difference between the two groups. At best the process may be moderated by slowing the rate at which the difference between the scores of the higher-scoring group and the scores of the lower-scoring group approaches the asymptotic level.

Test manufacturers have developed a statistical method for flagging pre-test items which have statistically significantly greater adverse impact upon an identifiable group of candidates. "Differential item functioning" (DIF) is a "statistical . . . property of a test item in which different groups of test takers who have the same total test score have different average item test scores or, in some cases, different rates of choosing various item options."[6] The DIF method stratifies members of the two groups into tiers of candidates who attained the same total test score. Candidates who have the same total test score are simply the candidates who have the same mean correct answer rate. The DIF method then compares the correct answer rates of members of the two groups within each tier of equal scoring candidates. The DIF method is impervious to the fact that one group of candidates may be predominately in the higher-scoring tiers and that the other group of candidates may be predominately in the lower-scoring tiers. In effect, the DIF method flags pre-test items on which the difference between the correct answer rates of the two groups is statistically significantly greater than the difference between the mean and correct answer rate of the two groups. The DIF method does not flag pre-test items which have correct answer rate differences which are greater, but only statistically insignificantly greater, than the mean correct answer rate difference. Even if the DIF method were rigorously applied and all flagged items were discarded, the effect would be merely a slight slowing of the process of approaching the asymptotic difference in group test performance. In practice, test manufacturers have not rigorously applied the DIF method. Standard 7.3 of the *Standards* states,

> "When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research results should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups."[7]

The *Standards* continue, "Although it may not be possible prior to first release of a test to study the question of differential item functioning for some such groups, continued operational use of a test may afford opportunities to check for differential item functioning."[8] The *Standards* abound in cautions concerning reliance upon the DIF method, for example,

---

6. *Id.* at 175.

7. *Id.* at 81.

8. *Id.* at 81-82.

> "When differential item functioning is detected, test developers try to identify plausible explanations for the differences, and then, they may replace or revise items that give rise to group differences if construct irrelevance is deemed likely. However, at this time, there has been little progress in discerning the cause or substantive themes that account for the differential item functioning on a group basis. Items for which the differential item functioning index is significant may constitute valid measures of an element of the intended domain and differ in no way from other items that show nonsignificant indexes. When the differential item functioning index is significant, the test taker must take care that any replacement items or item revisions do not compromise the test specifications."[9]

III.
THE BEST FITTING REGRESSION EQUATION

Admissions test scores and previous academic grades are both attractive candidates for inclusion in the admissions decision process. Admissions test scores and previous grade-point averages are both expressed quantitatively and, therefore, both amenable to mathematical and statistical evaluation. A linear regression equation is one method for combining the admissions test scores and the previous grade-point average into one composite predictor.

A linear regression equation is derived for each academic unit (e.g., Law School, Undergraduate College, etc.) to provide a predicted criterion value. For example, law schools derive an equation to predict first-year law school grade-point average (FYGPA). The equation assigns the weights to applicants' undergraduate grade-point averages (UGPA) and LSAT scores which best predict the applicants' future first-year law school grade-point averages. The equation is derived using a least-squares method. Obviously, deriving the equation requires that the applicants already have first-year law school grade-point averages. Therefore, the derivation is done retrospectively using the data from one or more previous admission cycles. The equation has the form

$$\text{predicted FYGPA} = (a)\,(\text{UGPA}) + (b)\,(\text{LSAT}) + (c).$$

The mathematical technique finds the values of a, b and c which produce the best fit to the actual FYGPA. The mathematical definition of the best fit is the solution, i.e., values of a, b and c, which minimizes the sum of the squared differences between each student's predicted FYGPA and actual FYGPA. This best-fitting least-squares solution may represent a very close fit or a very poor fit. The goodness-of-fit is expressed as the correlation between the predicted FYGPA and the actual FYGPA. Graphically the goodness-of-fit may be visualized on a scatter-plot which contains a point for each student. The horizontal axis of the graph represents the predicted FYGPA values and the vertical axis of the graph represents the actual FYGPA values. The point for an individual student is graphed at the intersection of the

9. *Id.* at 40.

student's predicted FYGPA and actual FYGPA. If the fit between predicted and actual FYGPA values is good, the scatter-plot looks like a swarm of points concentrated around an upward sloping line starting near the lower left-hand corner of the graph and culminating in the upper-right hand corner of the graph. The line around which the points are concentrated is called the regression line. (As used in the next section of this paper, the slope of the regression line represents the correlation between the predicted FYGPA and the actual FYGPA. If the regression line does not pass through the 0,0 point on the horizontal and vertical axes, i.e. the origin, then the point at which the regression line crosses the vertical axis is called the intercept.)

Admissions decisions, at least in part, are based upon the predicted first-year grade-point averages or their equivalents. (Institutions may construct a matrix of ranges of prior grade-point average and admissions test scores. Or, institutions may transform the predicted first-year grade-point average into another index. In either case, these methods are functionally equivalent to the predicted first-year grade-point average.) At various stages in the selection process admissions decisions may also be based upon additional applicant characteristics. An attack on any of the additional applicant characteristics will invariably lead to a comparison between the actual admissions decisions and the hypothetical admissions decisions which would have been made if selection had been based exclusively on the predicted first-year grade-point average. That is, the complainant is put in the position of having to, or wanting to, rely upon the predicted first-year grade-point average as the index of applicant qualifications. The reliance upon the predicted first-year grade-point average as the index of applicant qualification is bolstered by interpreting the correlation between the predicted and actual first-year grade-point average to be an indicator of the validity of the regression equation which generated the predicted first-year grade-point average. This presumption, of course, is predicated upon the acceptance of law school grades, or more specifically first-year law school grades, as the criterion for evaluating the validity of the admission process. Furthermore, the reliance upon the predicted first-year grade-point average is misplaced because the predictions are retrospective, performed only on the set of matriculating accepted applicants.

In practice, acceptance or denial of a particular student is not based solely on predicted first-year grade-point average or its equivalent. Not all applicants with predicted first-year grade-point averages, or their equivalents, above a strict cut-off value are accepted and not all applicants with predicted first-year grade-point averages, or their equivalents, below a strict cut-off value are denied. If other admissions criteria, whether explicit or implicit, objective or subjective, have any predictive power then the observed correlation between predicted and actual first-year grade-point averages is only partly attributable to the regression equation and partly attributable to the other admissions criteria. For example, in every admissions cycle, there are applicants with high predicted first-year grade-point averages who are denied admission because other information in the applicant's folder discourages reliance on the accuracy of the prediction. A true measure of the predictive validity of the regression equation is never available because the admission process is not amenable to experimental design. Matriculating accepted applicants are never a random sample of applicants or even a sample created by a strict cut-off value. There are no actual first-year grade-point average data for those denied applicants with high or low predicted first-year grade-point averages.

IV.
THE ACCURACY OF PREDICTED FIRST-YEAR GRADE-POINT AVERAGES

The statistical test of the accuracy of a predicted value, in this instance the predicted first-year average grade-point average, is the goodness-of-fit of the predicted values to the subsequent actual values. That is, the predicted first-year grade-point average of each matriculated student is compared one year later to the actual first-year grade-point average which the student attains.

The Law School Admission Council provides a Report on First-Year Performance under the general category of LSAT Correlation Studies.[10] The Fall 1997 report states the purpose of the report as, "Correlation studies provide useful information on the predictive validity of measures used in the admissions process. Law schools generally employ LSAT scores and undergraduate grade[-]point average (UGPA) to help predict academic success. This correlation study describes the relationships among first-year law school grade-point averages, LSAT, and UGPA for students who recently completed their first-year of law school and helps determine how well these measures are functioning as predictors. Where there is a strong and stable relationship between LSAT and UGPA and first-year averages (FYAs), LSAT and UGPA may be used to predict performance for future applicants."[11] The methodology is described in the following manner, "LSAT validity analysis is based on multiple-linear regression. Regression analysis yields a formula for predicting law school first-year average from a weighted combination of LSAT scored and UGPA. Certain statistics are presented to help evaluate the effectiveness of these predictors, both alone and in combination. The results of regression analyses are particular to the study group upon which they are based and may vary somewhat from one study year to the next as the population changes. To increase stability of the results, LSAT Correlation Studies are usually based upon the three most recent entering classes. The number of entering classes that can actually be included may be affected by changes in grading scale."[12]

The Law School Admission Council's Report on First-Year Performance does not include a calculation of the goodness-of-fit of the predicted first-year grade-point average made prior to admission with the actual first-year grade-point average maintained after matriculation. The Law School Admission Council's Report on First-Year Performance contains only a comparison of a newly derived predicted first-year grade-point average, made after admission base upon the subsequent actual first-year grade-point average, with the actual first-year grade-point average. The extent to which the correlation between this newly derived predicted first-year grade-point average and the actual first-year grade-point average "provide[s] useful information on the predictive ability of measures used in the admissions process"[13] is reduced by (1) the extent to which the mean of the first-year grade-point average

---

10. Law School Admission Council, *LSAT Correlation Studies: Rep. of First-Year Performance* (Fall 1997).

11. *Id.* at 3.

12. *Id.*

13. *Id.*

distribution changes from one year to the next, (2) the extent to which the variance of the first-year grade-point average distribution changes from one year to the next, (3) the extent to which the slope of the regression line changes from one year to the next, (4) the extent to which the intercept of the regression line changes from one year to the next, (5) the extent to which the departure from linearity of the best-fitting regression line changes from one year to the next, and (6) as explained in the last paragraph of the previous section of this paper, the unascertainable extent to which the goodness-of-fit is attributable to the other criteria used in the selection process.

V.
THE EFFECTS OF ABOLISHING AFFIRMATIVE ACTION

The demographic results of admissions selection processes traditionally have been monitored and adjusted to avoid undesired imbalances and to achieve desired class composition. The demographic characteristics which have often been considered are the representation of regions within a state or a country, the ratio of male and female applicants accepted, the racial or ethnic composition of the entering class, the alumni representation within the accepted students, the socioeconomic distribution of the accepted applicants, the sufficiency of athletic skills within the entering body, the ratio of public to private secondary school graduates within the entering student class, the percentage of students of a given religion among the accepted applicants, or the projected number of students who might select various academic majors. These considerations have had both exclusionary motives and inclusionary motives. Affirmative action as a means of increasing the admission of minority applicants has been adopted at many institutions, and now has been abolished at some institutions.

Recent experience at the law schools of the University of Texas, Austin, the University of California, Berkeley, and the University of California, Los Angeles has shown that the abolition of affirmative action results in a quantum drop in African-American enrollment to zero or near-zero levels. Large drops in enrollment of other minority groups also have occurred. Although alternative means of increasing minority enrollment have been discussed, none of the methods have been demonstrated to succeed.[14] In fact, none of the "new models" has been demonstrated to have any prospect of success.

I have attempted to explain how the test-item selection process inevitably creates an admissions test which irreversibly impacts on a lower-scoring distinguishable group of applicants. Additionally, I have attempted to explain how reliance on the regression equation which predicts first-year grade-point average as an index of student qualifications is misplaced. Barring the discontinuance of reliance on the regression equation for admissions decisions, it is reasonable to view affirmative action as the only available antidote to test development procedures which have reduced African-American enrollment to zero or near-zero levels and have substantially reduced the enrollment of other minority groups.

June 21, 2000

14. *See, e.g.*, Law School Admissions Council, *New Models to Assure Diversity, Fairness, and Appropriate Test Use in Law School Admissions* (Oct. 1999).