DOCUMENT RESUME

ED 378 202 TM 022 546

AUTHOR	Fan, Xitao; And Others
TITLE	Ethnic Group's Representation in Test Construction Sample and Test Bias.
PUB DATE	Aug 94
NOTE	33p.; Paper presented at the Annual Meeting of the American Psychological Association (102nd, Los Angeles, CA, August 12-16, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ethnic Groups; Genetics; High Schools; *High School Students; Models; Nature Nurture Controversy; Nonparametric Statistics; Psychometrics; *Sampling; Selection; Test Bias; *Test Construction; Test Format; Test Items
IDENTIFIERS	Texas; *Texas Assessment of Academic Skills

ABSTRACT

The hypothesis that faulty classical psychometric and sampling procedures in test construction could generate systematic bias against ethnic groups with smaller representation in the test construction sample was studied empirically. Two test construction models were developed: one with differential representation of ethnic groups (White, African American, Hispanic, and Asian) in each sample, and the other with exclusive representation of one ethnic group in each sample. Impact of group representation on the test item selection process and on ethnic groups' performance on the test forms developed was examined systematically using both parametric and nonparametric statistical techniques. Empirical findings based on the Texas Assessment of Academic Skills results over 190,000 11th graders consistently revealed that, for the tests constructed under the two models, there was no systematic bias against the group(s) with smaller representation in the test construction sample. Findings lead to the conclusion that the theory of genetic-environmental interaction, the theoretical underpinning of the hypothesis, may not be applicable in accounting for the dynamics of human testing. Seven tables present the data. (Contains 17 references.) (Author/SLD)

ETHNIC GROUP'S REPRESENTATION IN TEST CONSTRUCTION SAMPLE AND TEST BIAS

Xitao Fan

Utah State University

Victor L. Willson
Jerome T. Kapes

Texas A&M University

Paper presented at 102nd Annual Convention of American Psychological Association, Los Angeles, California, August 12-16, 1994. Session #: 3050.

ABSTRACT

The purpose of this study was to test empirically the tenability of the hypothesis that faulty classical psychometric and sampling procedures in test construction could generate systematic bias against ethnic groups with smaller representation in the test construction sample. Two test construction models were developed: one with differential representation of ethnic groups (White, African-American, Hispanic, and Asian) in each test construction sample, and the other with exclusive representation of one ethnic group in each test construction sample. The impact of group representation on the test item selection process and on the ethnic groups' performance on the test forms thus developed was examined systematically using both parametric and nonparametric statistical techniques. The empirical findings consistently revealed that, for the tests constructed under the two models, there was no systematic bias against the group(s) with smaller representation in the test construction samples. The findings lead to the conclusion that the theory of genetic-environmental interaction, which is the theoretical underpinnings of the hypothesis, may not be applicable in accounting for the dynamics of human testing.

Background

The issue of test bias has been both socially and emotionally sensitive and volatile for decades (Berk, 1982; Hilliard, 1984; Jensen, 1980; Reynolds and Brown, 1984). In the arena of psychological and educational measurement, nothing has invited more emotional debate and criticism than the mere mention of test bias, and nothing is as "vexing and thorny to test developers and test makers as the criticism that pertains to sex, racial, and ethnic test bias" (Berk, 1982. p.1). To the public, the question of test bias is often a highly emotional one, and test bias is often perceived as equal to unfairness, injustice, prejudice, or discrimination. To the professional circle of test makers and test users, potential test bias can be an important factor that may lower the quality of a test as a measuring instrument and produce systematically unreliable and erroneous results for certain client groups with regard to their abilities, aptitudes, or educational achievement. As a matter of fact, the concern over test bias has been so great that both legislation and litigation have become weapons in the debate over the controversy (Berk, 1982; Educational Testing Service, 1985; Green, 1981; Hickman and Reynolds, 1986-87; Mehrens and Popham, 1992; Reynolds and Brown, 1984.)

Traditionally, the focus of test bias discussion has been on the cultural aspect of test items, i.e., the test is biased because there is an excessive cultural loading on the test items which tends to favor the majority. As a result, the minority test-takers obtain lower test scores not because of any deficiency in knowledge or intelligence on their part, but because of the test-items' excessive cultural baggage which works against them. It has been rare to find arguments that question the validity of the classical psychometric procedures by which most of the tests are constructed.

Previously, the possibility was minimally discussed that ethnic group representation in the test development sa~ple might affect the item selection process (Green and Draper, 1972; Jensen, 1980). Jensen (1980), in his discussion of the "standardization fallacy", raised the question of how to take into account the minority groups in the standardization process. The standard practice in test development is population proportionate sampling, which means that the minority groups invariably have smaller proportions in test standardization sample, just as they do in the population. Jensen (1980) asked whether this sampling practice was sufficient for the claim that the minority groups had been properly taken into account to ensure the absence of test bias. It was tentatively suggested that it might be appropriate to carry out the item selection process separately for subgroups that were either equal in size or had large enough sample sizes to "allow for comparable statistical inferences regarding the psychometric properties of the test" (Jensen, 1980, p. 373). This reasoning recognized the potential that unequal representation of subgroups in the standardization sample might affect the item selection process in such a way so as to produce test bias.

Harrington (1975, 1984, 1988) challenged the integrity of classical psychometric procedures and practices by hypothesizing that the observed test performance differences among some ethnic groups might be the result of faulty psychometric procedures. Using an experimental model of animal testing, Harrington presented evidence that the classical psychometric procedures for item selection favored the genetic group(s) that had larger representation in the test construction sample. Generalizing from his animal testing model to human testing, Harrington stated that the sampling scheme and the item selection technique in classical psychometric theory might be inherently flawed, and it was this flawed theory and practice that were the prime suspects responsible for the observed mean score difference among certain ethnic groups. This hypothesis adds a new dimension to the debate of test bias.

The empirical evidence to support Harrington's hypothesis was from an animal testing model with three experiments, two of which are discussed here. Six genetically different groups of rats were involved in the animal testing model. The first experiment was designed to develop six nominally parallel test forms. The second experiment was designed to test if the six test forms developed in the first experiment were actually parallel.

In the first experiment, six independent test development samples were created. Each sample consisted of different proportions of the six genotypes of rats, with the proportions of the genotypes varied systematically within and across the samples. Maze items were administered to all the six test tryout samples and the items were scored. Based on the item statistics from each of the six test tryout samples, a fixed number of items were then selected to form six "parallel" test forms. The criterion for item selection was the highest item-total correlation, a criterion widely accepted for item selection in psychometric theory and practice (Crocker & Algina, 1986; Nunnally, 1978). Harrington's reasoning for the first experiment was that, if group representation in the test tryout sample did not systematically affect the item selection process as classical psychometric theory assumes, then the six test forms developed on the six independent test tryout samples should be regarded as parallel forms with only random sampling error. But this basic assumption in psychometric theory would be tested in the second experiment when the six nominally parallel forms were "administered" to independent and homogeneous groups of the six genotypes of rats, and their performance on the six forms were compared.

In the second experiment, the six "parallel" test forms were "administered" (re-scored, since these rats already ran all the maze items) to six independent and homogeneous groups of the same genotypes of rats as in Experiment 1. The performance of each genotype of rats on the six "parallel" forms of the test were computed and compared statistically. It was shown that, for each genotype of rats, its mean performance systematically varied on the six "parallel" forms and positively correlated with the proportion of that genotype in the test development samples in Experiment 1. This indicated that the forms constructed in Experiment 1 were not parallel as they were assumed to be, and the proportion of groups in the test development sample systematically impacted item selection process with the consequence that the test tended to favor the group(s) with larger proportion in the test development sample.

Based on the results from his animal-testing model, Harrington argued that under the strict experimental conditions, his experiment was a direct test of the classical psychometric methods. He thus challenged classical test theory and practice on several grounds.

First of all, Harrington argued, test items are systematically, as opposed to randomly, selected into a test mainly by the criterion of internal consistency. This selected sample of test items cannot be said to be a random sample from the universe of all possible items, though they have to be treated as such for the purpose of score interpretation. But the necessary condition for making inferences or generalizations from a test score to a person's trait or ability is that the test items are a random sample from the universe of all possible test items. In the terms of analysis of variance as Fisher described it (Fisher, 1935), if a model had fixed effects (as opposed to random effects), no generalizations beyond the fixed effects could be made.

Second, the standard sampling procedure for the development of all major tests employs population proportionate sampling with regard to different ethnic groups. This sampling scheme entails that minority groups have smaller proportions in the test construction sample, which will result in their minimal impact on the item selection process.

Third, Harrington argued, the item selection process based on the internal consistency criterion has a strong tendency of favoring the group(s) with larger representation in the test construction sample. The impact which a group has on the item selection process is directly related to its proportion in the test construction sample. Due to this differential impact, the test will tend to be biased against minority groups.

Fourth, since the group membership has been a factor affecting the item selection process in the standard psychometric practice, and consequently inherent bias against minority group(s) has been built into the test, it is illogical to use the test to show that the groups are different. For the same reason, it is also inappropriate to use the end product of bias to show that bias does not exist.. This is especially so when we consider the possibility that this type of bias may affect all the test items within the same test. It also may affect both the criteria and predictor tests, since they are usually developed using the same faulty procedures, thus having the same systematic bias built into them.

The theoretical rationale offered by Harrington to account for the phenomenon observed in his study is termed genetic-environment interaction. Put simply, genetic-environment interaction theory means that genetically different organisms possess different response systems to different environmental circumstances. As a result, genetically different organisms may respond to the same environmental stimuli in different ways. The phenomenon of genetic-environment interaction exhibits itself most clearly in the process of natural selection. Harrington (1984, p. 130) has the following to say with regard to this:

> The theory of evolution posits that environmental circumstances tend to select out those genotypes best adapted to survival in the specific environment and that, in different environmental circumstances, different genotypes are selected. The theory is that different genes fit different environments better. This theory exactly defines the term genetic-environmental interaction. Different genotypes respond differently in different environments (emphasis original).

Harrington reasoned that, in the situation of test development involving different racial groups, genetic-environmental interaction would exhibit itself because the test items represented different environmental stimuli and different genotypes would have their respective optimal responses to different test items. The process of systematic item selection would tend to select those items on which the genetic group(s) with larger representation in the test construction sample had optimal responses, since those items would tend to appear as having better item-total correlation. On the other hand, the items to which minority group(s) had optimal responses would tend to be discarded from the item pool for not being consistent with the items on which the majority group had better performance, and statistically, not having good item-total correlations. As a result, the test items in the final item pool would tend to have systematic bias favoring the group(s) with larger representation in the original test development sample.

Harrington was the first to rationalize the issue using the genetic-environment interaction theory, thus providing a new perspective for the issue. Harrington's experiment and the results, though controversial and not widely known or accepted, are certainly intriguing and thought provoking. The implications of his hypothesis, if proved to be tenable, can be both theoretically and practically far-reaching. In response to some serious doubts which have been expressed with regard to the appropriateness of Harrington's generalization from his animal testing model to human testing (Jensen, 1984), Harrington (1988, p. 406) stated:

> The primary thrust of these studies stands independently of any questions of intelligence or intelligence testing. It is quite irrelevant whether or not animal intelligence is a model of human intelligence, whether maze performance is a measure of intelligence, or whether or not intelligence is *g*. The essential manipulations, ... were to control genotype and to carry out standard psychometric procedures for test construction. ... The data show simply that standard psychometric procedures under conditions of tight experimental control lead to two forms of minority test bias when groups differ genetically.

Besides the genetic factor considered in his study, Harrington also believed that the phenomenon observed in his experiment could also be expected with environmental differentiation for the groups (Harrington, 1975).

Since Harrington put forth his hypothesis, very few attempts have been made among measurement professionals to test the tenability of Harrington's hypothesis for human testing. Up to this date, to the author's knowledge, only two published studies are specifically related to the issue raised by Harrington: the study by Green and Draper (1972) and the study by Hickman and Reynolds (1986-87). The lack of attention Harrington's experiment had received from psychometric professionals caused some psychologists (Hirsh & Tully, 1982) to question the reluctance of the psychometric community to acknowledge the implications of Harrington's experimental model. Hirsh and Tully (1982) further challenged the psychometric community to test the hypothesis by evaluating similar experimental design for human testing.

Independent of Harrington's experiment, Green and Draper (1972) conducted an empirical study in a similar direction and used human achievement test data of different intact socio-ethnic groups. Their results were somewhat ambiguous and inconclusive. Hickman and Reynolds (1986-87) recognized the serious challenge posed by Harrington's hypothesis to classical test theory and the standard test development practice.

> "Harrington's hypothesis strikes at e very core of psychometric methods. If he is correct, true score theory, as we know it, would be devastated in its application to the study of individual differences" (p.

143).

Using human intelligence testing data, Hickman and Reynolds (1986-87) compared two test forms constructed on white and black samples respectively. Their results did not lend support to Harrington's hypothesis.

Up to this time, the two studies described above are the only two which examined the issue, and the results are far from being conclusive. The present study is designed as a direct empirical test for the hypothesis put forth by Harrington. More specifically, the following two research questions were to be addressed empirically in this study:

1) Is the test performance of an ethnic group related to the group's proportion in the original test development sample?

2) Do ethnic groups systematically perform better on test forms constructed on samples like themselves than they do on forms constructed on other ethnic groups?

Data Source and Methods

Data Source

A large scale testing program database of the Texas Assessment of Academic Skills (TAAS) was used for the study. TAAS is a criterion-referenced test battery with three subtests: Reading, Mathematics and Writing. In this study, only data from the Reading and Mathematics subtests were used, and both of them consisted of multiple-choice items. Due to the nature of criterion-referenced testing, TAAS tests are based on instructional content areas, and as such, TAAS test score validity is essentially content-based. The items in Reading and Math sections were selected into their respective item pools primarily based on how well the items matched the prespecified instructional objectives to be assessed in TAAS. Additionally, item difficulty as estimated by one-parameter IRT Rasch model was also considered mainly for the purpose of test equating. In the item selection process, no special efforts were made to select items so as to maximize internal consistency reliability, though items with very poor item-total correlations were usually scrutinized for possible defect (E. N. Morgan, Texas Education Agency, personal communication, May 1993).

The TAAS data used in this study were from the 1992 October administration of TAAS exit level tests normally taken by Grade 11 students. The items from both Reading (48 items) and Mathematics (60 items) were used. The subject pool contained over 190,000 subjects, with the breakdown for the four ethnic groups in the subject pool as in Table 1.

Insert Table 1 about here

Experiments and Replications

Two experiments were designed for the study, each with its own independent and exact replication on new samples. The replications of the experiments were designed to reduce the likelihood of chance discovery, and to assess the replicability of experimental results. To guarantee the independence of replications, the full data set was randomly and evenly split into two data sets before any samples were drawn from the data pool. These two halves of the original data set were used for the experiments and their replications respectively. Altogether, 8,000 subjects (2,000 for each ethnic group) were used in the two experiments and their independent replicatioıs.

Two tightly controlled experiments were conducted and each experiment had two stages: test construction and test administration. The first experiment used differential representation of ethnic groups in test development samples: four independent test development samples were created, with the proportions of four ethnic groups (White, Black, Hispanic and Asian) systematically varied within and across the samples (0%, 10%, 30%, and 60%), as represented in Table 2. Based on the four test development samples, four "parallel" test forms (Form 1 to Form 4) were constructed by selecting 50% of the test items from the original item pool, using item-total correlation as the sole item selection criterion.

Insert Table 2 about here

In the second stage of the experiment, all the four nominally parallel test forms were "administered" to each of four new and ethnically homogeneous test-taking samples (N=200 each).

In other words, each new and ethnically homogeneous test-taking sample took all the four test forms (Form 1 through Form 4), and its performance was compared across the forms. Since, in the data base, all subjects had already attempted all the test items, this "test administration" was essentially a statistical rescoring process: separately scoring only those items selected into each of the four "half" test forms (Form 1 through Form 4). If Harrington's hypothesis is correct, we should expect the performance pattern to emerge that, for any ethnic group, its performance on the four "half" test forms should tend to covary positively with that group's proportion in the original test development samples as in Table 1. To take Asian group as an example, the performance of this group's test-taking sample should tend to have the following performance pattern, since this group's proportion in the test development sample was largest for Form 4 (60%), and smallest for Form 1 (0%):

Form 4 > Form 3 > Form 2 > Form 1

Similar patterns should be expected for other ethnic groups, and the nonparametric Page's *L* statistic (Page, 1963) was used to test such an ordered hypothesis. Page's *L* statistic is a non-parametric statistic used for data of an ordinal nature. As a non-parametric statistical technique, no assumption of normal distribution is necessary, nor is the assumption of equal distances between treatments. It is specifically designed for testing the null hypothesis of equality of several means against an ordered alternative hypothesis (Page, 1963):

$$H_0: m_1 = m_2 = \cdots = m_k$$
$$H_1: m_1 > m_2 > \cdots > m_k$$

In using Page's _L_ statistic, it is required that the logic for the predicted order is theoretically reasonable. The rationale for such predicted order is certainly provided by Harrington's reasoning which was described in detail previously.

The second experiment used maximum representation of each ethnic group in test construction samples, as represented in Table 3. In other words, each of the four test construction samples was composed of 100% of one ethnic group only.

Again, four nominally parallel test forms were constructed based on the four independent samples as in Table 3 by selecting 50% of the items from the original item pool, using item-total correlation as the selection criterion. The four test forms (White form, Black form, Hispanic form, and Asian form) were then "administered" to four _new_ and _ethnically homogeneous_ test-taking sample (N=200 for each), with each test-taking sample taking all the four test forms. If Harrington's hypothesis is correct, it would be expected that, for each test-taking sample, its performance would be better on the form developed on sample like itself than on forms developed on samples of other ethnic groups. More specifically, according to Harrington's reasoning, the following performance patterns could be expected for the four test-taking samples in this experiment: the White group scores higher on White Form than on the other three forms; the Black group scores higher on Black Form than on the other three forms, etc. This expected performance pattern is displayed in Table 4. Since this experiment used the most extreme of unequal representation of ethnic groups in test development samples, it would be expected that the phenomenon as observed by Harrington, if it is true at all for human testing, would have maximum chance to exhibit itself.

Planned contrasts within repeated measure analysis of variance were conducted for individual ethnic groups to determine if the performance pattern for each of the ethnic groups conformed to the expectations under Harrington's hypothesis. As an example, for the White group, the contrast was set up such that their performance on Form 1 was compared with, and was expected to be higher than, the average of all the other three forms. Similar contrasts were set up for the other three test-taking groups. If the observed performance pattern of the ethnic groups consistently conformed to the expectations under Harrington's hypothesis, the results would be interpreted as providing support for Harrington's hypothesis; otherwise, the opposite conclusion would be drawn.

Results and Discussions

The first research question examined the possibility that the test performance of an ethnic group would positively covary with that group's proportional representation in the original test construction sample. Harrington originally put forth his hypothesis BY presenting animal testing data to demonstrate that this was the case. So in this sense, the answer to this question is the direct test of the tenability of Harrington's hypothesis or the appropriateness of Harrington's generalization from animal testing model to human testing situations.

Table 5 and Table 6 present the performance ranks for the four test forms for each of the four independent and ethnically homogeneous test-taking samples, the calculated Page's *L* statistic, and the probability associated with the Page's *L* statistic. The maximum *L* statistic (if data conformed to the predicted order perfectly), and the minimum *L* statistic (if data behaved just exactly the opposite of the predicted order) are also presented for easy reference.

The two tables of Page's *L* statistics show that the order effect as predicted by Harrington's hypothesis did not appear in either of the "half-test" forms of TAAS tests, neither did it appear in the replications. The actual performance rank order of each of the test-taking samples on the four test forms consistently showed only random order, rather the systematic order as predicted by Harrington's hypothesis.

Insert Table 5 about here

Insert Table 6 about here

To facilitate understanding of the lack of statistical significance of the Page's *L* statistic, and to show the absence of the order effect predicted by Harrington's hypothesis, the mean performance ranks of the ethnic groups were plotted against the group proportion in the test construction samples for the four test forms. Figures 1 to 5 present these plots, with the plot based on Harrington's data (1984) being presented first for easy comparison. It is obvious that the order effect as exhibited by Harrington's data did not appear in any of the plots generated in this study.

Insert Figure 1 to Figure 5 about here

For the second experiments and its independent replication, as discussed previously, contrasts could be set up so that, for any test-taking ethnic group sample, its mean on the test form constructed on a sample like itself could be contrasted with, and expected to be higher than, its grand mean on the three other test forms. Table 7 presents the results of such analysis.

The contrast analysis showed that, out of the 16 contrasts, only six conformed to the expectation of Harrington's hypothesis, a number smaller than 50-50 random chance occurrence. Furthermore, in all instances, the difference between the contrast means was very small. The contrast analysis clearly indicated the absence of the performance patterns as predicted by Harrington's hypothesis.

Empirical results from both experiments (differential group representation and maximum group representation) and their independent replications did not show any indication that the phenomenon predicted by Harrington's hypothesis was true for human testing. These results cast serious doubts on the tenability of Harrington's genetic-environment interaction reasoning for human testing, and on the validity of his generalization from his animal testing model to the human testing situation. In short, in both the experiments and their replications, the empirical results did not indicate any systematic test bias caused by classical psychometric procedures and practices, as Harrington's reasoning predicted. Consistently, for any ethnic group, even being 100% represented in the test construction sample failed to provide any clear advantage to that group.

Though the genetic-environment interaction theory as applied to natural selection of organisms is generally accepted, the appropriateness of its application to human aptitude testing has not been adequately addressed. This study attempted to test empirically the viability of Harrington's genetic-environment interaction theory as applied to human testing situation. The consistent lack of support for Harrington's hypothesis indicates that the genetic-environment interaction theory may not be appropriate for accounting for the dynamics of human testing.

Considering the utmost importance of replicability of research results in any scientific endeavor, one strength of the present study lies in its well-designed replications to avoid chance discovery. In this sense, this feature of replication adds considerable weight to the validity of the conclusions drawn from this study.

Although it is not entirely clear how and why the animal testing model as in Harrington's experiment differs from the human testing model in the present study, the discrepancy of results may be tentatively accounted for from two perspectives: the difference in genetic homogeneity between animal testing and human testing, and the difference in terms of surviving values associated with test performance in animal testing model versus in human testing situation.

In Harrington's animal testing model, animals of pure genetic strains were the subjects. Obviously, for the human testing situation, the social and ethnic groups are very heterogeneous genetically. Harrington (1988) recognized this difference between animal testing and human testing and discussed that it might be more difficult to detect the effects as described from his animal testing model.

Another more important difference between animal testing and human testing may be related to the function of testing for genetic selection. The necessary condition for applying genetic-environment interaction theory to account for the dynamics of human testing is to assume that different test items in human IQ or achievement testing are comparable to the different "environments" in biological sense. Biological genetic-environment interaction theory as evidenced in natural selection of organisms obviously assumes that the responses to the environments have direct survival values, and the natural selection is carried out according to the principle of "survival of the fittest". Theoretically, the appropriateness of generalizing the results from Harrington's animal testing model to human testing situation may be questioned from the perspective of surviving value difference between animal testing and the human testing. In the animal testing model, the maze items have life-sustaining function for the animals, and as a result, they are related to genetic selection. In this sense, the genetic-environment interaction theory may be appropriate to account for certain aspects of the animal testing. In the human testing situation, on the other hand, performance on aptitude tests may not be related to genetic selection, or the tasks may be devoid of daily life-support functions. The absence of the life-sustaining function of human testing and the absence of any direct surviving values associated with aptitude test items for human testing may have rendered the biological genetic-environment interaction theory inappropriate for accounting for the dynamics of human testing.

REFERENCES

Berk, R. A. (Ed.). (1982). Introduction. In R. A. Berk, (Ed.), Handbook of methods for detecting test item bias. Baltimore: The Johns Hopkins University Press.

Crocker, L., and Algina, J. (1986). Introduction to classical and modern test theory. San Francisco: Holt, Rinehart and Winston, Inc.

Educational Testing Service. (1985). Legal issues in testing. ERIC Clearinghouse on Tests, Measurement, and Evaluation. Princeton, New Jersey: Educational Testing Service. (ERIC No. ED289884).

Fisher, R. A. (1935). The design of experiments. London: Oliver and Boyd.

Green, D. R. and Draper, J. F. (1972). Exploratory studies of bias in achievement tests. Paper presented at the Annual Meeting of the American Psychological Association. September, 1972. Honolulu, Hawaii. (ERIC No. 070794).

Harrington, G. M. (1975). Intelligence tests may favor the majority groups in a population. Nature, 258, 708-709.

Harrington, G. M. (1984). An experimental model of bias in mental testing. In C.R. Reynolds and R.T. Brown (Eds.), Perspectives on bias in mental testing. New York: Plenum.

Harrington, G. M. (1988). Two forms of minority-group test bias as psychometric artifacts with an animal model (Rattus norvegius). Journal of Comparative Psychology, 102, 400-407.

Hickman, J. A., and Reynolds, C.R. (1986-87). Are race differences in mental test scores an artifact of psychometric methods? A test of Harrington's experimental model. The Journal of Special Education, 20, 409-430.

Hilliard, A. G. (1984). IQ Testing as the emperor's clothes: A critique of Jensen's Bias in Mental Testing. In C. R. Reynolds and R. T. Brown (Eds.), Perspectives in mental testing, New York: Plenum Press.

Hirsch, J., and Tully, T. P. (1982). The challenge is unmet. The Behavioral and Brain Sciences, 5, 324-327.

Jensen, A. R. (1980). Bias in mental testing. New York: The Free Press.

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C.

R. Reynolds and R. T. Brown (Eds.), Perspectives on bias in mental testing. New York: Plenum Press.

Mehrens, W.A. and Popham, W.J. (1992). How to evaluate the legal defensibility of high-stakes tests. Applied Measurement in Education, 5, 265-283.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. Journal of the American Statistical Association, 58, 216-230.

Reynolds, C. R., and Brown, R. T. (Eds.). (1984). Bias in testing: Introduction to the issues. In C.R. Reynolds and R.T. Brown (Eds.), Perspectives on bias in mental testing. New York: Plenum Press.

Table 1: Ethnic Composition of the Subject Pool

Ethnic Group	Freq.	%	Cumulative Freq.
White	98166	52.05	98166
Black	24714	13.10	122880
Hispanic	59918	31.77	182798
Asian	5815	3.08	188613

Table 2: Sampling Design for Experiment #1: Differential Representation of the Four Ethnic Groups

Test Construction Samples (N)	% of Representation in the Samples				Resultant Test Forms
	60%	30%	10%	0%	
Sample 1 (300)	White	Black	Hispanic	Asian	Form 1
Sample 2 (300)	Black	Hispanic	Asian	White	Form 2
Sample 3 (300)	Hispanic	Asian	White	Black	Form 3
Sample 4 (300)	Asian	White	Black	Hispanic	Form 4

Table 3: Sampling Plan for Experiment #2: Maximum Representation of Ethnic Groups

Test Construction Samples		Ethnic Groups				Resultant Test Forms
Sample	N	White	Black	Hispanic	Asian	
1	300	100%	0%	0%	0%	White Form
2	300	0%	100%	0%	0%	Black From
3	300	0%	0%	100%	0%	Hispan. Form
4	300	0%	0%	0%	100%	Asian Form

Table 4: Expected Performance Pattern in Experiment #2 for Four Homogeneous Test-Taking Samples on Four Test Forms

	Test Forms			
Test-Taking Samples (N)	White (100% White)	Black (100% Balck)	Hispanic (100% Hispanic)	Asian (100% Asian)
White (200)	High	Low	Low	Low
Black (200)	Low	High	Low	Low
Hispanic (200)	Low	Low	High	Low
Asian (200)	Low	Low	Low	High

Table 5: Testing Order Effect for Test-Taking Samples - TAAS Reading

Proportion in Tryout Sample	60%	30%	10%	0%	Page's L[a]	P for Page's $L_{Calc.}$
Expected Rank Order	1	2	3	4		
		Experiment				
White	4[b]	3	1	2		
Black	2	4	3	1		
Hispanic	1	2	4	3		
Asian	3	2	1	4		
					$L_{Calc.}=99$	$p>.05$
		Replication				
White	4	1	2	3		
Black	4	2	1	3		
Hispanic	3	4	1	2		
Asian	2	1	3	4		
					$L_{Calc.}=98$	$p>.05$

[a] $L_{max}=120$; $L_{min}=68$; $L_{crit.}=111$

[b] Performance rank order among four "parallel" tests for each ethnically homogeneous test-taking sample

Table 6: Testing Order Effect for Test-Taking Samples - TAAS Math

Proportion in Tryout Sample	60%	30%	10%	0%	Page's L[a]	P for Page's $L_{Calc.}$
Expected Rank Order	(1)	(2)	(3)	(4)		
		Experiment				
White	3[b]	4	2	1		
Black	1	4	3	2		
Hispanic	2	1	4	3		
Asian	2	3	1	4		
					L_{Calc}=102	$p>.05$
		Replication				
White	4	4	1	2		
Black	2	3	4	1		
Hispanic	1	2	3	4		
Asian	4	1	2	3		
					$L_{Calc.}$=100	$p>.05$

[a] L_{max}=120; L_{min}=68; $L_{crit.}$=111

[b] performance rank order among four "parallel" tests for each ethnically homogeneous test-taking sample

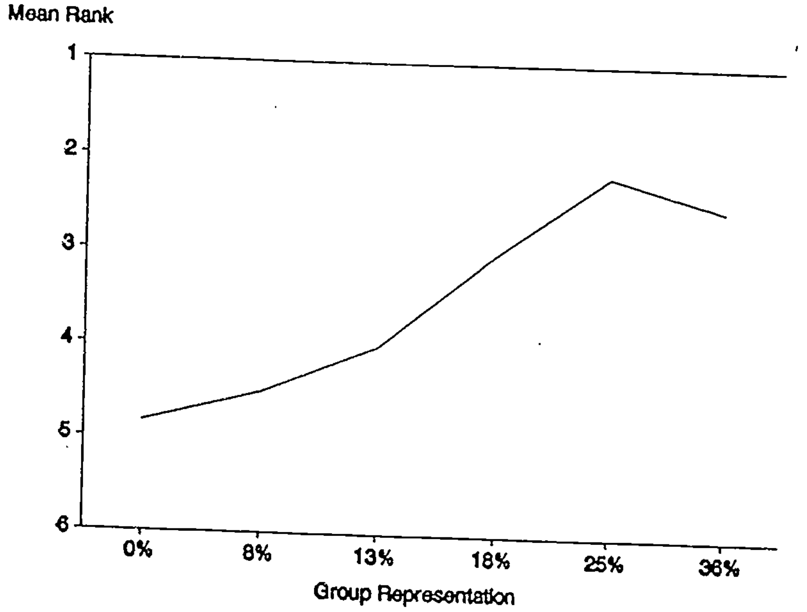

Figure 1: Harrington's Data Showing Order Effect

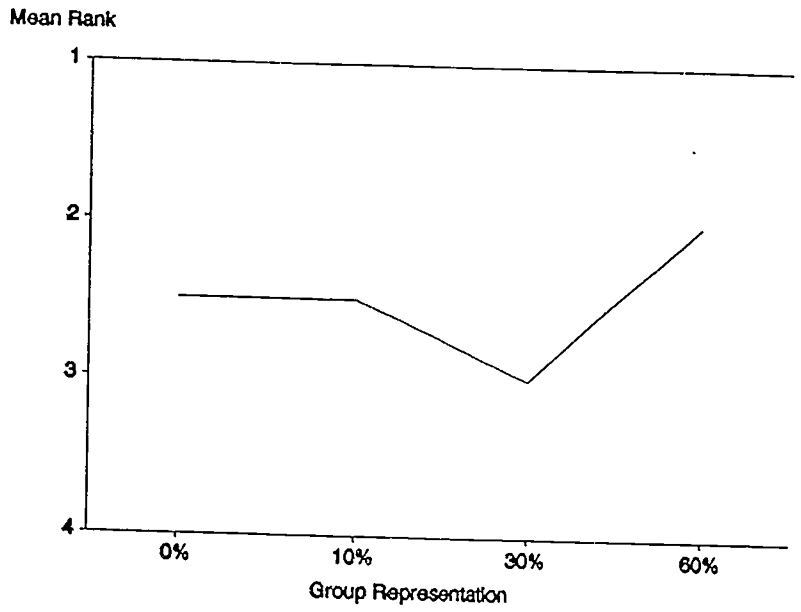

Figure 2: Mean Rank and Group Representation - Math Experiment #1

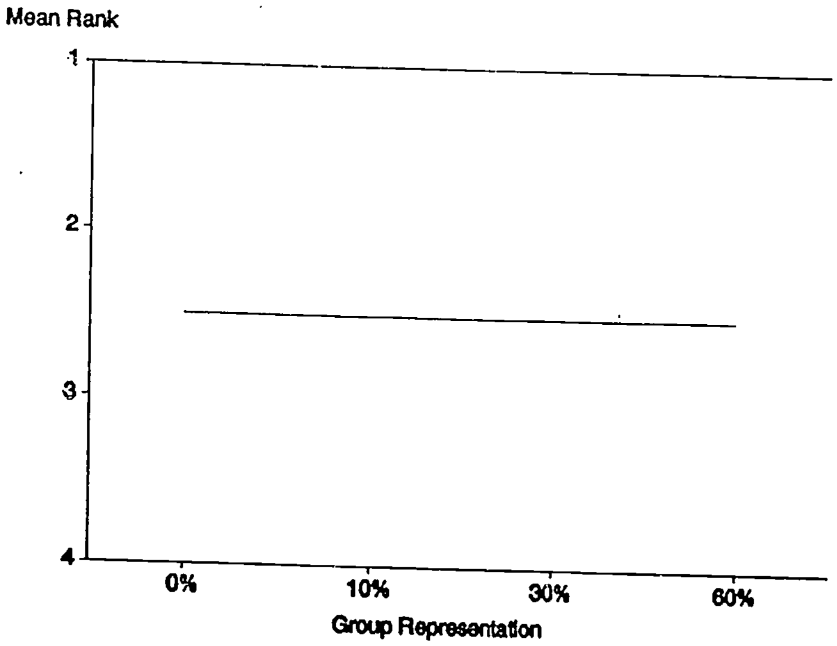

Figure 3: Mean Rank and Group Representation - Math Replication of Experiment #1

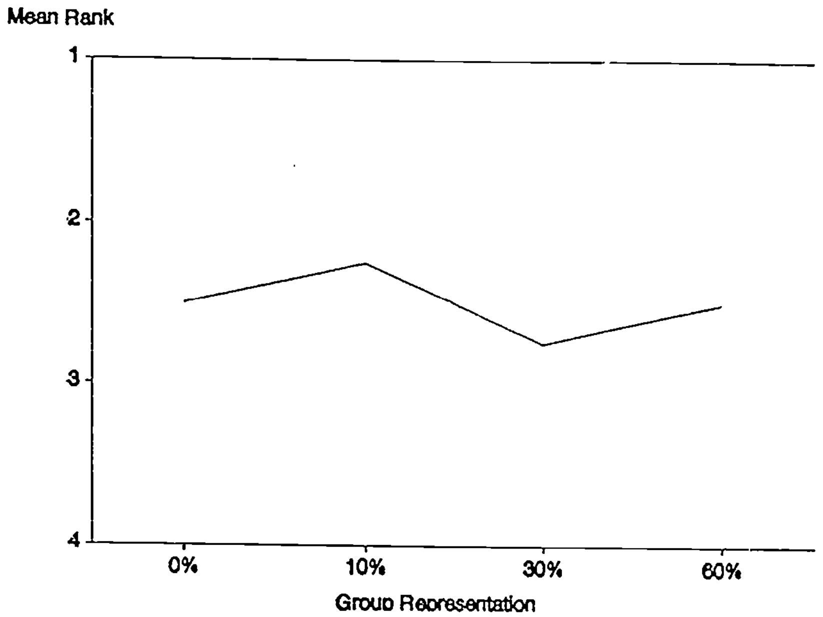

Figure 4: Mean Rank and Group Representation - Reading Experiment #1

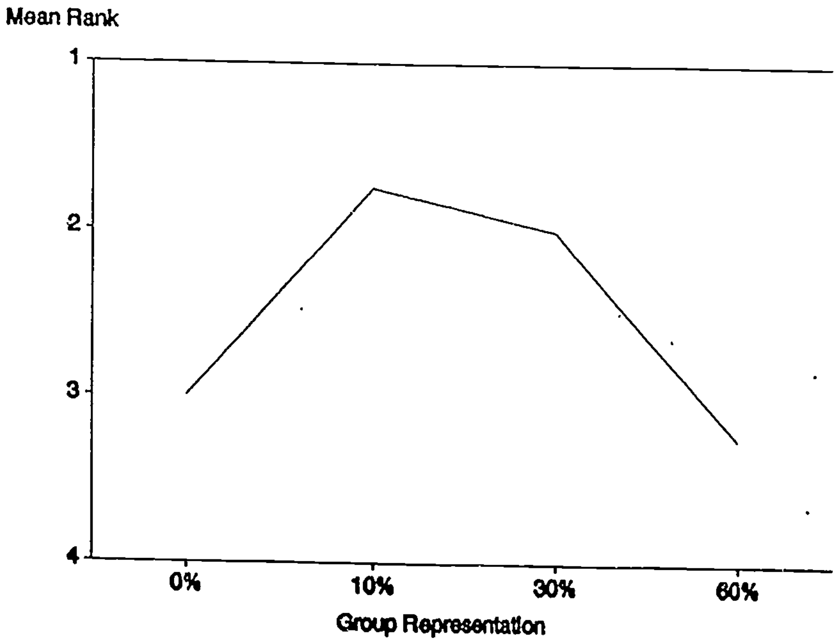

Figure 3: Mean Rank and Group Representation - Reading Replication of Experiment #1

Table 7: Results of Contrast Analysis for Experiment #2 and its Replication

Reading

	Mean		If Direction Agrees with Harrington [a]	In Favor of Harrington Hypothesis
Test-Taking Ethnic Group	Own Form	Other Three Forms	Contrast P Value	
White	19.08 [b]	20.33		no
	(19.69)	(20.56)		no
Black	17.67	17.02	.001	yes
	(16.76)	(16.92)		no
Hispanic	17.50	17.61		no
	(17.25)	(16.75)	.001	yes
Asian	19.40	19.08	.001	yes
	(18.49)	(18.22)	.007	yes
Math				
White	22.81	23.11		no
	(22.95)	(23.22)		no
Black	17.29	17.16	.16	no
	(17.01)	(16.74)	.06	no
Hispanic	19.32	18.85	.001	yes
	(19.00)	(18.66)	.001	yes
Asian	24.52	25.24		no
	(23.68)	(24.58)		no

[a] Statistical tests were presented only for those contrasts in which the difference direction conforms to Harrington's hypothesis.

[b] The mean from Experiment, and that from Replication is in parenthesis below.