Journal of Educational Statistics
Summer 1984, Volume 9, Number 2, pp. 93–128

ACCOUNTING FOR STATISTICAL ARTIFACTS IN ITEM BIAS RESEARCH

LORRIE SHEPARD
University of Colorado

GREGORY CAMILLI
Human Systems Institute

and

DAVID M. WILLIAMS
University of Colorado

KEY WORDS. *Item bias, test bias, IRT (latent trait) applications.*

ABSTRACT. Theoretically preferred IRT bias detection procedures were applied to both a mathematics achievement and vocabulary test. The data were from black and white seniors on the High School and Beyond data files. To account for statistical artifacts, each analysis was repeated on randomly equivalent samples of blacks and whites (n's = 1,500). Furthermore, to establish a baseline for judging bias indices that might be attributable only to sampling fluctuations, bias analyses were conducted comparing randomly selected groups of whites. To assess the effect of mean group differences on the appearance of bias, pseudo-ethnic groups were created, that is, samples of whites were selected to simulate the average black-white difference.

The validity and sensitivity of the IRT bias indices was supported by several findings. A relatively large number of items (10 of 29) on the math test were found to be consistently biased; they were replicated in parallel analyses. The bias indices were substantially smaller in white-white analyses. Furthermore, the indices (with the possible exception of χ^2) did not find bias in the pseudo-ethnic comparison. The pattern of between-study correlations showed high consistency for parallel ethnic analyses where bias was plausibly present. Also, the indices met the discriminant validity test—the correlations were low between conditions where bias should not be present. For the math test, where a substantial number of items appeared biased, the results were interpretable. Verbal math problems were systematically more difficult for blacks.

Overall, the sums-of-squares statistics (weighted by the inverse of the variance errors) were judged to be the best indices for quantifying ICC differences between groups. Not only were these statistics the most consistent in detecting bias in the ethnic comparisons, but they also intercorrelated the least in situations of no bias.

Researchers (see Bond, 1981; Cole, 1981) have taken three general approaches to research on the question of test bias: (a) predictive validity studies in selection situations; (b) investigation of external biasing factors such as the race of examiner, test-wiseness of examinees, and speed of administration;

and (c) construct or content validity studies of the internal structure of the test. The present research is focused on test item-bias methods, which are subsumed in the last category of inquiry. Item-bias methods are statistical procedures intended to test whether items function equivalently in two groups. Therefore, they address the basic validity question: Does the test (or individual items in the test) measure what it purports to measure for both groups?

There are numerous item-bias methods (see Berk, 1982; Rudner, Getson, & Knight, 1980a; Shepard, 1981). Most rely on an item-by-group interaction criterion of bias; that is, statistical adjustments are made for overall group differences, and then items that are relatively more difficult for one group are flagged as potentially biased.

A standard operating assumption should be discussed regarding item-bias techniques. Because they lack an external criterion, they can only be used to detect relative, not pervasive, bias in a test (Petersen, 1977). The various methods either use total test score (or estimated abilities from the total set of items), or average *p*-value differences to define the "typical" difference between groups; this then becomes the standard of "unbiasedness" against which individual items are compared. Thus, if there is bias in the determination of this typical group difference, it will go undetected by these techniques. Despite this limitation, it has been argued that item-bias procedures may be the preferred approach for understanding the nature of bias and for uncovering irrelevant difficulties in items whose meanings change for members of different groups (Shepard, 1982). The predictive validity models of test fairness involve an external criterion but are not without fault. Petersen and Novick (1976) demonstrated that the various models for defining equal regressions (i.e., equal predictive validity for two groups) are mutually contradictory. Moreover, Linn (1982) has recently explained how differential measurement error for two groups could obscure predictive bias. Finally, there is the actuarial problem (Shepard, 1982). Predictive validity studies look only at the magnitude of the correlation between test and criterion; they do not distinguish between relevant and irrelevant sources of relationship. Nor do they examine whether the combination of predictors that maximize the correlation are equally defensible. Test-item bias methods let us look more directly at what we are measuring. They leave for a second step the question of how measures of separate traits should be combined to make selection or other test-based decisions.

Among item-bias techniques, the theoretically preferred method is the three-parameter item response theory (IRT) or item-characteristic-curve (ICC) method. It is preferred because of its sample invariant properties that make it less likely that true group differences will be mistaken for bias. Hunter (1975) and Lord (1977) have demonstrated heuristically that bias techniques based on classical test theory (such as *p*-value differences or point-biserials) will produce invalid indices of bias in the presence of group mean differences. Because *p* differences interact with item discrimination, items that are merely more discriminating (i.e., better measures of the trait in both groups) will have bigger differences in performance. Furthermore, the variability of a particular group and how "centered" an item is for that group will artifactually control the item's discriminating power. Methods such as empirical ICCs (Green & Draper, 1972) and chi-square procedures (Scheuneman, 1979) were intended to be approximations to the IRT method. These procedures are crude, however, and will still confound real group differences with bias because of regression effects. The one-parameter latent-trait method (or Rasch model) shares the theoretically sample invariant properties of the three-parameter model. However, the Rasch model is not recommended for bias detection because it will confound other sources of model misfit (particularly differences in item discrimination) with item "bias" (see Divgi, 1981; Ironson, 1982; Shepard, Camilli, & Averill, 1981).

The conceptual definition of bias using the three-parameter IRT model and specific procedures will be explained in the Method section. The three-parameter IRT method was applied in this study because it is theoretically the most sound. Its superiority is relative, however, rather than absolute; bias detection using IRT is not without problems. First there are estimation problems due to sampling fluctuations. Even with reasonably large sample sizes, as in this study, it is possible that misestimated item parameters for separate groups could create or obscure item-characteristic curve differences when the two groups are compared. More important, there may be larger sources of error when samples are very different. Even the theoretical claims for the model are said to be true only when the model holds. The following discussion is focused on the potential for obtaining invalid bias indices, even with IRT methods. First, however, a digression is in order regarding substantive intepretation of bias indices. Difficulties encountered when trying to make substantive interpretations of bias analyses may be linked to the problem of statistical artifacts.

Substantive Interpretations of Bias

Given the increasing concern over cultural bias in tests, a strong impetus to the development of statistical screening techniques was the apparent failure of judgmental methods for identifying biased test questions. That is, even minority experts, sensitive to the issue of cultural loading in test questions, could not predict with better than chance success what type of items would be difficult for members of particular groups (Jensen, 1977; Plake, 1980; Sandoval & Miille, 1980). For example, Jensen (1976) found that an item on the WISC often cited for its dependence on white middle-class values, "What is the thing to do if a fellow (girl) much smaller than yourself starts to fight with you?" was actually easier for blacks to answer. If biased test questions were not obvious to expert judges, then perhaps statistical detection procedures could uncover more subtle changes in the meaning of items for different groups.

A more disappointing result—after numerous statistical bias studies—has been that here too expert judges are often at a loss to explain the source of bias in items with large bias indices. For instance, in an early study, Lord (1977) found that 46 of 85 items on the verbal SAT were significantly different for blacks and whites (bias was sometimes against whites). But, in studying the items identified as biased, no particular insights could be gained to explain the differential performance. It was hoped that the use of statistical bias techniques would lead to substantive generalizations about the nature of items found to be biased against specific groups. For example, Scheuneman (1979) found that negatively worded items were biased against blacks. This type of consistent finding turned out to be more the exception than the rule. Raju (in Green et al., 1982) described the serious problems faced by test publishers who may decide to discard statistically deviant items even though they are unable to explain why they are biased "in terms of the content." The disconcertingly large number of uninterpretable statistically-biased items leaves the test maker with a dilemma. Has the statistical indicator uncovered a real instance of bias, revealing a blind spot in the conceptualization of the test construct, or is the large bias index a statistical artifact, that is, not a valid sign of bias? (see Shepard, 1981). We are aware of the potential for artifactual errors in the bias methods. These artifactual explanations become all the more plausible when the bias results seem uninterpretable.

Control of Statistical Artifacts

There are both random and systematic sources of error associated with IRT bias indices. For example, because the current statistical theory for maximum likelihood estimation in item response theory is only approximate, conclusions regarding group differences may be sample dependent (Bougon & Lissak, 1981; Lord, 1980). In fact, Lord (1980) proposed that replication or reliability studies should be carried out on independent but randomly equivalent groups of blacks and whites. One purpose of this research is to conduct such stability comparisons.

There is also some art involved in implementing IRT procedures. Choices made in applying computer programs to arrive at maximum likelihood estimates can have small but important effects. In the use of IRT specifically for studying item bias, a difficult stage in the procedures is the equating phase. Parameters must be estimated separately for two groups but then equated to the same scale for comparative purposes. Errors in the equating can produce spurious instances of bias. In simulation studies, for example, three-parameter indices of bias only correlated on the order of .80 with generated bias (see Merz & Grossen, 1979; Rudner, Getson, & Knight, 1980b). Because the three-parameter logistic model was initially used to create bias in these data, near perfect correlations might have been expected between simulated and detected bias. One must conclude that either sampling fluctuations or some implementation problem, as suggested above, prevented better "recovery" of the bias that had been built in.

In addition to the replication or cross-validation method to control for unreliability, the degree of error in bias indices also can be assessed by baseline studies. Lord (1980) created random groups which he called "reds" and "blues" to check on the number of "significantly" biased items in a condition where there should be no bias. Similarly, Ironson and Subkoviak (1979) used white-white comparison groups to assess the validity of both classical and latent-trait bias indices. In this research we will use white-white comparison groups to study not only the amount of bias due only to sampling errors but also to establish numeric baseline values for interpreting bias indices that lack distribution theory.

Artifactual problems associated with random sampling error will be exacerbated when the groups to be compared differ in mean ability on the test. Angoff (1982) suggested that the classical *p*-value method would be more valid for detecting bias (rather than confounding differential difficulty with item discrimination) if groups were equal or nearly equal on the trait initially. For example, we might expect fewer artifactual problems in most male-female comparisons than with black-white comparisons. Even the three-parameter indices, which are theoretically sample invariant, may be unstable when differences between groups are large. We know, for example, that latent-trait equating procedures are more stable for horizontal equatings (different tests, same grade) than for vertical equatings (same test, different grades). The vertical equating problem, where groups are located in very different regions of the ability continuum, is analogous to bias studies where groups have large mean group differences. Again referring to classical methods, Angoff (1982) suggested that an appropriate baseline for interpreting bias indices would not be just randomly equivalent white groups but white groups that differed in mean ability. This is the same analysis strategy used by Jensen (1974) when he created pseudo-ethnic groups; that is, white groups selected on age (with a mean difference of two years) to simulate the average black-white differences. When we know that the statistical techniques are intended to correct for group differences but may do so imperfectly, the point is to simulate with all-white data what the effect might be of mean differences only. In the current research, pseudo-ethnic comparisons are used in addition to randomly equivalent white groups.

Purpose Summary

The substantive purpose of this research is to study item bias between black and white examinees on both a mathematics and vocabulary test. The theoretically preferred three-parameter IRT approach will be used with optimal techniques for computing bias indices based on previous research. The major focus of the research is methodological rather than substantive. To assess the *amount* of artifactual (i.e., spurious) bias identified, both randomly equivalent white groups and extreme white groups (pseudo-ethnic comparisons) will be used. To identify the particular instances of unstable bias indices, cross-validations or replication analyses will be performed with randomly equivalent black and white groups.[1] Finally, items found to be consistently biased will be inspected for substantive characteristics. It is hypothesized that once artifactual instances of bias are better controlled, the results should be more interpretable than they have been in previous bias studies.

Method

Data Source

The data used for this investigation are from the High School and Beyond (HSB) data files available from the National Center for Education Statistics. The HSB sample includes over 30,000 high school sophomores and 28,000 seniors, from a representative probability sample of the nation's tenth- and twelfth-grade populations. The test and questionnaire data were collected in the spring of 1980 by the National Opinion Research Center under contract with the National Center for Educational Statistics. The particular examinees selected for study were black and white seniors. Unless otherwise specified (e.g, when pseudo-ethnic groups were created), the subsamples used were selected at random from the larger group of 3,377 black or 17,928 white seniors (excluding Hispanics). The following study samples were created[2]:

Math Test

Comparison 1: $W1$, $B1$	n = 1,500 whites, 1,500 blacks
Comparison 2: $W2$, $B2$	n = 1,500 whites, 1,500 blacks

[1] The design employed in this study is characterized by two replications of the black-white contrasts. Although two replications are better than one instance of the comparison, we realize it is an arbitrary decision to stop at two. A number of techniques are suitable for repeating the comparison from a fixed number of observations including jackknifing and bootstrapping (see Efron, 1982). However, because one step in any resampling procedure must be the execution of IRT analyses, repeated sampling experiments using IRT software would be prohibitively expensive.

[2] In keeping with the constraints of LOGIST, examinees were excluded if their scores on the relevant test were 0% or 100%. Abilities (θ's) cannot be accurately estimated for subjects who are at the ceiling of the test; zero scores may not represent a valid administration and surely do not indicate that the examinee has been "measured."

(Sampling for comparison 1 and comparison 2 was without replacement, so the samples were independent.)

Comparison 3: *W*1, *W*2	white samples from comparison 1 and comparison 2
Comparison 4: *B*1, *B*2	black samples from comparison 1 and comparison 2
Comparison 5: *W*1, and Pseudo B (*W*3)	white sample from comparison 1 and white sample (n = 1,500) selected to match the distribution of *B*1 on math total score

Vocabulary Test

Comparison 1: *W*4, *B*4	n = 1,500 whites, 1,500 blacks
Comparison 2: *W*5, *B*5	n = 1,500 whites, 1,500 blacks
Comparison 3: *W*4, *W*5	white samples from comparison 4 and comparison 5

The two tests analyzed were the senior mathematics and vocabulary tests. Although both tests were administered in two parts, we treated the combined item sets (32 in math and 27 on vocabulary) as single tests. More will be said about the nature of the items and the factor structure of the combined tests in the discussion of results. The math test was primarily a basic skills test involving simple operations, reading a graph, calculating a per unit cost, and comparing rates or distances. Four of the math items required some familiarity with basic algebra or geometry at a level that is usually included in K–8 curricula. The vocabulary test was more difficult; the average percentage of seniors answering items correctly was .46 compared to .58 for math items. The content of the vocabulary test was clearly aimed at a higher grade level, either high school or, in some cases, college level. The words are almost exclusively Latin roots. According to word frequency counts in representative materials for grades one to nine (Carroll, Davies, & Richman, 1971), the vocabulary items are relatively unfamiliar. The standardized frequency indices indicate that the words are found either not at all in junior high level materials or at a rate of about one in a million or one in ten million words. Examples of words with similar frequencies would be: fanatic, recalcitrant, marauding, permeated, reciprocating, and crevasses.

IRT Bias Method

IRT permits the expression of examinee responses to individual test items as a function of the underlying ability or trait measured by the test. In the Results section of the paper, these item characteristic curves (ICCs) or item response functions are illustrated. The horizontal dimension of the graph is the ability (or θ) scale. For each item, a monotonic increasing curve reflects the probability of getting the item correct for increasing values of θ. The ICC is defined by three parameters: (a) the *a* parameter is proportional to the slope of the curve at the inflection point and represents the item's discrimination; (b) the *b* parameter reflects the item's difficulty and is a location on the θ ability dimension (when there is no guessing, *b* is the point where the probability of getting the item correct is 50%); and (c) the *c* parameter is often referred to as the "pseudo-guessing" parameter (it is the lower asymptote of the curve and represents the probability of getting the item correct for examinees of extremely low ability).

The IRT method for detecting item bias is based on the comparison of item-characteristic curves estimated separately for two groups. The ICCs reflect the probability of getting the item right as a function of ability. If an item is unbiased, examinees of equal ability should have equal probabilities of success on the item regardless of group membership; that is, the ICCs for different groups should be the same. If ICCs for two groups differ by more than sampling error, the item is apparently not measuring the same underlying trait for both groups (at least not to the same degree) and is therefore "biased."

It should be noted that IRT models rest on an assumption of unidimensionality; the items in the test all assess the same underlying trait, and ability on that trait only, not some other trait, influences item performance. In the Results section we present factor analyses as supporting evidence that this assumption is met for these data. We did not devote much attention to prior testing of this assumption apart from routine factor analyses, however, because in a sense the bias studies themselves are addressed to the issue of unidimensionality. In fact, multidimensionality will be detected as bias by the IRT method so long as group differences are not uniform across the different traits. As stated by Linn, Levine, Hastings, and Wardrop (1981), "Bias may generally be conceptualized as multidimensionality confounding differences on a primary trait with differences on a secondary trait" (p. 161).

The LOGIST program (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976) was used to estimate the person abilities and item parameters. Because the chance (*c*) parameters are difficult to estimate even with large sample sizes, we followed the technique suggested by Lord (1980), whereby *c*'s were estimated in a combined analysis and then fixed at that value for the separate analyses within ethnic groups. This aggregate or composite analysis was done only at the level of each study comparison. That is, black and white samples chosen for separate replications were not combined to get even more stable estimates of *c*. Rather, we wished to preserve the separateness of each comparison study and do each as if it were the only data available to the researcher. Additional particular information about how the LOGIST program was implemented is given in the Results section.

Scale Equating

Once item parameters (defining the ICCs) have been estimated separately for two ethnic groups, the ICCs for each item must be compared to detect bias. However, because the θ scales from an IRT analysis are arbitrary (set with $\bar{X} = 0$ and $s = 1$ for the given sample), the ICCs from separate analyses are not directly comparable. The ICCs must first be equated to the same scale. To make the adjustment, we used a linear transformation of the b parameters as described in Linn, Levine, Hastings, and Wardrop (1980), Appendix B). Briefly, the equating is determined by a best fitting line that adjusts for the difference in average item b values and has a slope equal to the ratio of the standard deviations of the two sets of b's. In computing means and variances, b parameters were weighted by the inverse of the variance error in estimating b. Therefore, items with poorly estimated b's contributed least to the equating. Once the linear equation was obtained, the b parameters for the second sample were recomputed in the metric of the first group. In this case, the black parameters were converted to the white scale.

After the b's were adjusted, the same equating constants (the slope and intercept) were also used to transform the θ values. Finally, the a parameters were equated, using the inverse of the slope determined for the b equating (Lord, 1980, p. 36). The c parameters do not require equating.

Bias Indices

For an individual test item, bias is defined as the difference in the probability of answering correctly, given equal ability. Once item characteristic curves have been adjusted to the same scale, differences in the probability of a correct response are synonymous with differences in the ICCs. Several different indices were used to quantify ICC differences between groups.

Unsigned Indices

Unsigned area (UA). As described in Shepard, Camilli, and Averill (1981), the area between two ICC functions was evaluated as a definite integral for an item i:

$$\int_{-3}^{+3} \{\hat{P}_{iW}(\theta) - \hat{P}_{iB}(\theta)\} d\theta,$$

where $\hat{P}_{iW}(\theta)$ and $\hat{P}_{iB}(\theta)$ are the estimated probabilities of a correct response given θ for the white and black groups, respectively.

Sum of squares 1 (SOS1). Linn, Levine, Hastings, and Wardrop (1980) developed both weighted and unweighted sums of squares statistics. The following index is similar but is "self-weighting" in that squared differences in probabilities are summed for every value of θ that *occurs,* rather than creating intervals on the θ scale and using the midpoint of each interval. Thus, probability differences in the region where the most data occur will contribute more to the index.

$$\mathrm{SOS1}_i = \frac{1}{n_W + n_B} \sum_{j=1}^{n_W + n_B} \{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\}^2.$$

The j subscript counts all instances of θ for either group ($n_W + n_B$). When θ_j is an obtained value in the white group, the probability difference is computed as if the value had also been observed in the black group and vice versa. n_W and n_B are the numbers in the white and black groups, respectively.

Sum of squares 2 (SOS2). SOS2 is similar to SOS1 except that squared differences in probabilities were weighted by the inverse of the variance error of the difference in ICCs for each given value of θ.

$$\mathrm{SOS2}_i = \frac{1}{n_W + n_B} \sum_{j=1}^{n_W + n_B} \frac{\{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\}^2}{\hat{\sigma}^2_{P_{iW} - P_{iB}}},$$

where $\hat{\sigma}^2_{P_{iW} - P_{iB}}$ is the variance error of the difference in estimated probabilities and other terms are as defined above. Formulae for computing the variance error of a point on an estimated ICC are given in Linn, Levine, Hastings, and Wardrop (1980, Appendix A). Following their reasoning, P differences contributed less to the weighted index if either P were poorly estimated.

Chi-square (χ^2). Lord (1980) proposed an asymptotic significance test to compare a and b differences between groups simultaneously. By the following chi-square formula, the hypothesis is tested that the vector of a and b differences is different from the vector (0, 0):

$$\chi_i^2 = \underset{\sim}{V}_i{}' \, \hat{\underline{\Sigma}}_{ab_i}^{-1} \, \underset{\sim}{V}_i,$$

where

$$\underset{\sim}{V}_i = \begin{Bmatrix} \hat{a}_{iW} - \hat{a}_{iB} \\ \hat{b}_{iW} - \hat{b}_{iB} \end{Bmatrix}.$$

Signed Indices

All of the unsigned indices reflect the magnitude of the differences between ICCs for two groups, but they do not carry signs to indicate the direction of the bias, that is, which group has the lower probability of a correct answer. In fact, when the item-characteristic curves cross, one group is not consistently disadvantaged. Rather, one group is ahead in one region of the graph, but behind in another region.

Signed indices are computed similarly to the corresponding unsigned indices. When the ICCs do not cross, the absolute values of the indices are the same but with a sign attached to show the direction of bias. When the curves cross, "bias" in two regions of the curve may be offsetting.

Signed area (SA). When the ICCs for two groups did not cross in the region from -3 to $+3$, the SA was equal to the UA except that a negative sign was attached if the item was biased against whites, if whites had a lower probability of getting the item right given θ. If the ICCs did cross, θ^* was found as the root of the equation $P_W(\theta) = P_B(\theta)$. Then the integral was evaluated from -3 to θ^* and θ^* to $+3$. The signed area was the difference between these two areas and carried the sign of the larger area.

Sum of squares 3 (SOS3). SOS3 is the "signed sum of squares" index analogous to SOS1. By multiplying $[\hat{P}_{iW}(\theta) - \hat{P}_{iB}(\theta)]$ times its absolute value, rather than squaring the difference, the sign of the difference is preserved.

$$\text{SOS3}_i = \frac{1}{n_W + n_B} \sum_{j=1}^{n_W+n_B} \{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\} \,|\, \hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j) \,|\, ,$$

where terms are as defined previously.

Sum of squares 4 (SOS4). SOS4 is the weighted sum parallel to SOS2. It is computationally the same as SOS3 except that every squared difference is weighted by the inverse of the variance error of the difference.

$$\text{SOS4}_i = \frac{1}{n_W + n_B} \sum_{j=1}^{n_W+n_B} \frac{\{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\} \,|\, \hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j) \,|}{\hat{\sigma}^2_{P_{iW} - P_{iB}}} ,$$

where terms are as defined previously.

a and b differences (AD), (BD). Simple differences between *a* parameters $(a_W - a_B)$ and *b* parameters $(b_W - b_B)$ were computed.[3] These differences were of interest because of their relation to other bias indices. The *a* and *b* differences were not interpreted as bias indices themselves because separately they do not characterize ICC differences well. As Linn et al. (1980) have shown, *a* and *b* parameters could be substantially different for two groups but not result in any practical differences in the ICCs. This would be true, for example, if an item were difficult in both groups so the *b*'s diverged in a θ region where few examinees existed.

Results and Discussion

Factor Analyses

Factor analyses were performed on the mathematics and vocabulary tests to determine if they could be considered unidimensional. Tetrachoric correlations were obtained using the total senior sample of 25,069. Principal factors were extracted after iterating for communalities. Each factor with an eigen-

[3] Note that the black *b* was always subtracted from the white *b*, because this corresponds to the order of subtraction in all the other indices. However, because a high *b* means the reverse of a high probability of correct response, the signs will have the opposite meaning.

value greater than one was retained for rotation. An oblique solution was obtained by direct oblimin transformation with $\Delta = 0$ (Harman, 1967).

In the math test, the first unrotated factor accounted for 30% of the total variance. Four additional factors that met the minimum eigenvalue criterion of one accounted for 5%, 4%, and 3% of the variance, respectively. Similarly, on the vocabulary test, the first unrotated factor comprised 30% of the total variance. In this case, there were only two additional factors, accounting for 6% and 4% of the variance, respectively. For both tests we interpreted the results as reasonably strong evidence of unidimensionality. First, the percentage of total variance explained by the first factor exceeds Reckase's (1979) minimum of 20% needed to assure stable item parameters. Also, an inspection of the scree plot of latent roots suggested that only the first eigenvalues deviated from the gradual rise that could be expected from factoring uncorrelated variables.

As stated previously, the singularity of the trait measured by the test for the population generally may not be measured so well in particular subgroups. It is the purpose of the bias analysis to test whether this unidimensionality is true for both black and white groups.

Bias Indices with Replication for the Math Test

In Table I, the bias indices are reported for mathematics items across five comparisons. To simplify the amount of information, a reduced set of indices is presented. The signed and unsigned areas and the asymptotic chi-square have been the most popular in the past. As we will see from the correlational results, the weighted and unweighted SOS statistics are highly similar, but there is some evidence for preferring the "behavior" of the weighted versions; therefore only SOS2 and SOS4 are shown.

The first two sets of indices, from comparison 1 and comparison 2, are the *replicated bias studies* based on randomly equivalent groups of blacks and whites. As will be discussed in the next section, a baseline for judging the magnitude of the bias indices was obtained from the white-white analysis. Index values that exceeded the largest number occurring in the white-white analysis are starred as biased in Table I.

A substantial number of items with ICC differences replicated across studies. Out of the 29 math items, for which ICCs were estimated in both groups, 10 were consistently biased. (Three of these items were biased in favor of blacks.) We said items were consistently biased if they exceeded the cutoff on most or all of the indices in *both* studies. It is worth noting that fully one-third of the test items appear to be deviant by this relatively stringent rule. When we caution that item-bias methods are internal methods and hence unable to detect constant bias, this does not imply that we are limited to finding only one or two discrepant items.

TABLE I

Signed and Unsigned Bias Indices for Math Items in Five Comparison Studies

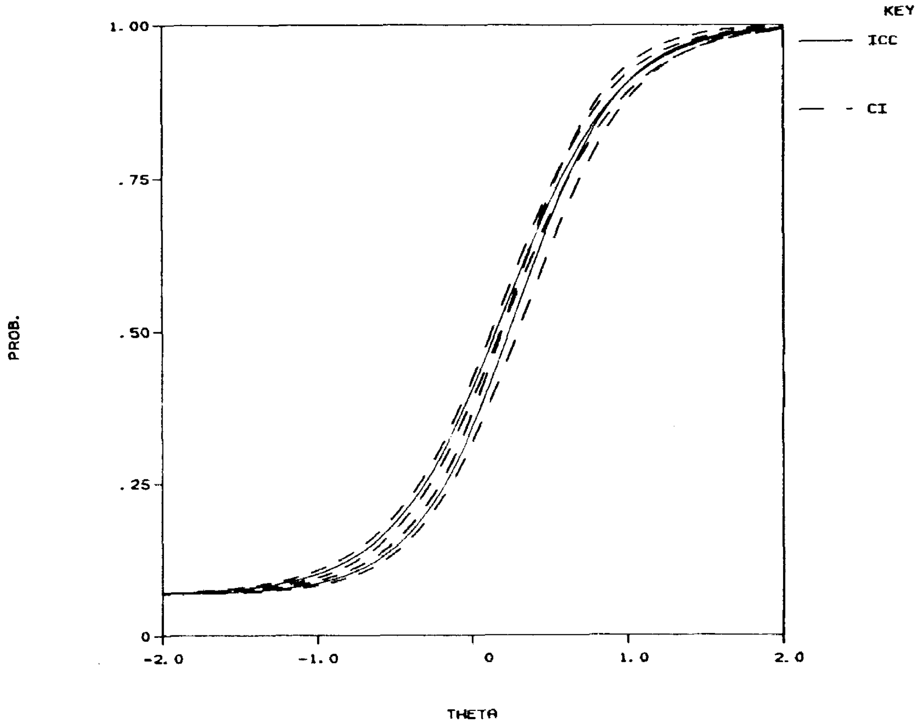
	Comparison 1: W1, B1					Comparison 2: W2, B2					Content Classification: verbal or numeric	Comparison 3: W1, W2					Comparison 4: B1, B2					Comparison 5: W1, Pseudo B (W3)				
	Unsigned			Signed		Unsigned			Signed			Unsigned			Signed		Unsigned			Signed		Unsigned			Signed	
Item	UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4		UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4	UA	SOS2	χ^2	SA	SOS4
1	.23*	4.86	49.19*	.08	-4.76	.05	3.69	6.40*	-.01	-3.68	n	.03	.04	.15	.03	-.04	.24*	38.49*	20.53*	-.03	38.31*	.05	1.13	2.16	-.05	-1.13
2	.13	2.35	3.76	.08	2.34	.18	3.23	4.73	.01	3.14	v+n	.05	.36	.34	.00	-.36	.13	.70	2.03	-.13	-.70	.15	1.50	3.39	.15	1.50
3	.33*	14.60*	20.93*	.33*	14.60*	.46*	17.82*	21.63*	.29*	17.43*	v	.17	.81	2.45	.12	-.40	.17	.93	1.81	.00	-.65	.31*	2.38	11.35*	.26*	1.86
4	.24*	6.37	10.94*	.24*	6.37	.33*	16.16*	14.11*	.23*	16.11*	n	.01	.02	.04	-.01	-.02	.08	.33	.69	-.07	-.33	.13	3.16	3.68	.11	3.16
5	.03	.77	.95	-.01	.50	.13	2.68	9.24*	.13	2.68	n	.09	1.22	4.71	-.08	.14	.03	.89	2.04	.03	.89	.09	1.59	12.07*	-.09	-1.55
6	.42*	11.42*	17.63*	.12	9.99*	.53*	14.14*	20.30*	.29*	13.04*	v	.07	.17	.71	.07	.17	.15	1.31	3.34	.14	1.31	.10	1.13	1.21	.03	1.09
7	.25*	8.84*	19.81*	.25*	8.84*	.23*	14.08*	22.50*	.23*	14.08*	v	.11	4.35	2.67	.03	-4.23	.08	.82	1.78	-.01	.26	.12	1.10	4.29	.12	1.10
8											n						.20	.71	3.65	.18*	-.40					
9	.19	1.68	5.69*	-.14	-1.45	.43*	7.51	16.74*	-.10	1.88	v	.12	1.67	2.14	.02	-1.35	.10	.51	1.31	.04	.37	.03	.11	.20	.03	.11
10	.14	4.16	6.66*	.01	.41	.14	2.84	4.81	.01	.86	n	.04	.26	.48	-.01	.03						.08	.99	2.40	.07	.98
11	.31*	15.48*	14.87*	.10	15.24*	.51*	30.25*	19.25*	.05	29.06*	v	.06	.40	.38	-.01	.36	.08	2.68	1.01	-.03	2.67	.14	4.59	5.04	.12	4.59
12	.21	21.11*	23.20*	-.11	-20.36*	.26*	31.60*	42.45*	-.22*	-31.46*	n	.06	.68	.72	.02	-.57	.06	1.52	2.96	-.06	-1.52	.13	2.65	3.47	.03	-1.67
13	.13	3.95	8.16*	.05	.07						n	.05	1.27	1.90	.05	1.26	.05	1.28	2.65	-.05	-1.28	.09	1.23	3.36	.06	.69
14	.46*	10.79*	33.55*	.29*	4.90	.50*	10.39*	33.69*	.34*	8.61	v	.13	2.69	2.47	.02	2.64	.11	1.45	3.37	.05	-.69	.16	2.16	5.68*	.16	2.16
15	.15	.40	2.27	.14	.26	.18	1.58	3.15	.13	-1.17	v+n	.12	2.42	4.22	-.04	1.57	.03	.57	.17	.00	-.57	.19	1.36	2.59	.13	-.91
16	.23*	16.78*	42.44*	-.23*	-16.78*	.19	8.02	20.88*	-.19*	-8.02	n	.06	1.16	2.32	-.04	-.98						.10	2.65	7.79*	-.10	-2.65
17	.23*	13.21*	23.38*	-.22*	-13.19*	.34*	24.40*	34.56*	-.27*	-24.19*	n	.10	2.83	3.70	.07	2.67	.05	.23	1.07	.05	.23	.16	7.25	6.46*	.04	6.31
18	.38*	6.93	13.40*	-.09	2.78	.27*	2.43	7.29*	-.18*	1.61	v	.09	.77	1.03	.03	.71	.07	1.98	1.49	-.02	-1.98	.32*	7.32	9.18*	.04	6.59
19	.01	.01	.02	.00	-.01	.02	.04	.08	-.02	-.04	n	.06	.63	1.25	-.06	-.63	.05	.24	.87	-.05	-.24	.15	2.06	4.13	.13	2.04
20	.14	3.40	7.08*	.14	3.40	.15	2.96	7.97*	.15	2.96	v	.06	.99	1.88	-.06	-.99	.08	.87	2.20	-.05	.47	.11	2.34	1.46	.01	-1.03
21	.08	3.61	5.11	.07	3.59	.07	2.09	2.42	.05	2.06	n	.03	.29	.71	.03	.29	.05	4.98	1.78	-.01	4.95	.06	1.27	2.49	.06	1.27
22	.23*	14.08*	29.20*	.23*	14.08*	.25*	13.92*	25.73*	.25*	13.92*	v	.04	.42	.92	.02	.32	.03	.64	.28	.02	-.63	.05	.55	1.47	-.04	-.44
23											v											.38*	2.60	5.25	.38*	2.60
24	.06	1.17	1.15	.03	1.16	.04	.28	.33	.03	.28	v	.10	.72	1.77	.05	-.30	.07	1.20	.72	.02	-1.19	.15	1.32	2.75	-.09	.65
25	.06	.89	.93	-.06	-.89	.19	.12	2.04	.19*	.07	n															
26	.06	.36	1.04	.04	-.04	.08	1.76	3.06	-.08	-1.76	n	.09	2.25	3.15	.02	1.38	.13	4.34	3.95	-.08	4.17	.05	.22	.85	.05	.20
27	.08	.24	1.73	.08	.09	.13	.41	3.09	.12	.11	n	.01	.02	.12	-.01	-.01	.03	.63	.53	.03	.63	.09	4.23	3.61	.02	-4.08
28	.08	1.35	3.90	-.03	-1.35	.06	1.69	1.83	-.03	1.27	n						.11	7.70	11.54*	.06	6.56	.07	.76	1.83	-.05	.29
29	.13	.61	2.73	-.13	-.61	.15	1.34	2.89	.06	1.32	n	.17	.59	4.57	-.17*	-.58	.26*	.73	4.21	-.16	.36	.01	.01	.02	.00	.01
30	.68*	28.60*	8.09*	-.26*	28.49*	.10	.38	.47	.06	-.38	v	.23*	8.72*	5.37*	.07	8.71*	.23*	2.93	2.03	.08	-2.93					
31	.04	.29	.31	.03	-.26	.05	.83	.50	.02	-.80	n											.03	.23	.38	.02	-.21
32	.06	1.44	.85	.00	-1.43	.09	.80	.56	.06	-.79	n	.01	.08	.08	-.01	.07	.14	.95	1.56	.13	-.94	.15	12.93*	1.46	.01	12.88*

Note: To establish a baseline for judging the magnitude of the bias indices, the largest value for each index from the white-white comparison was used. Indices that exceeded this value in other comparisons are starred as "biased." For the sake of consistency, the largest W1, W2 χ^2 of 5.37 was used; however, 5.99 is the critical value for statistical significance at α = .05.

Figures 1–4 are item characteristic curves for blacks and whites on several illustrative math items. Figure 1a shows the graph for an unbiased item. The two solid lines reflect the probability of a correct response, given θ, for blacks and whites. For all values of θ, the two groups have essentially equal probabilities of answering correctly. Figure 1b is an example of a biased item. The white curve is consistently above the black curve, so whites and blacks of equal ability do not have the same probability of success. (The curves for item 7 were similarly discrepant in comparison 2 with a slightly larger effect.)

The items in Figures 2a and 2b are also consistently biased in both comparisons. These graphs are more typical of most of the biased items in that the ICCs for the groups cross within the θ region of −2 to +2. Therefore, the bias in one region of the curve is partially offset by a reverse bias at the other end of the θ scale. Signed indices allow this cancelling effect to occur and therefore only show a large bias index if one group is overall more disadvantaged than the other. Even between the signed indices there is a difference in how bias is quantified. The signed area (SA) is a simple measure of the amount of squared difference between the curves. The signed SOS4 index is more heavily

FIGURE 1a. Comparison of white and black item-characteristic curves for study 1, item 21 on the math test. (Example of an item found unbiased in both comparisons.)

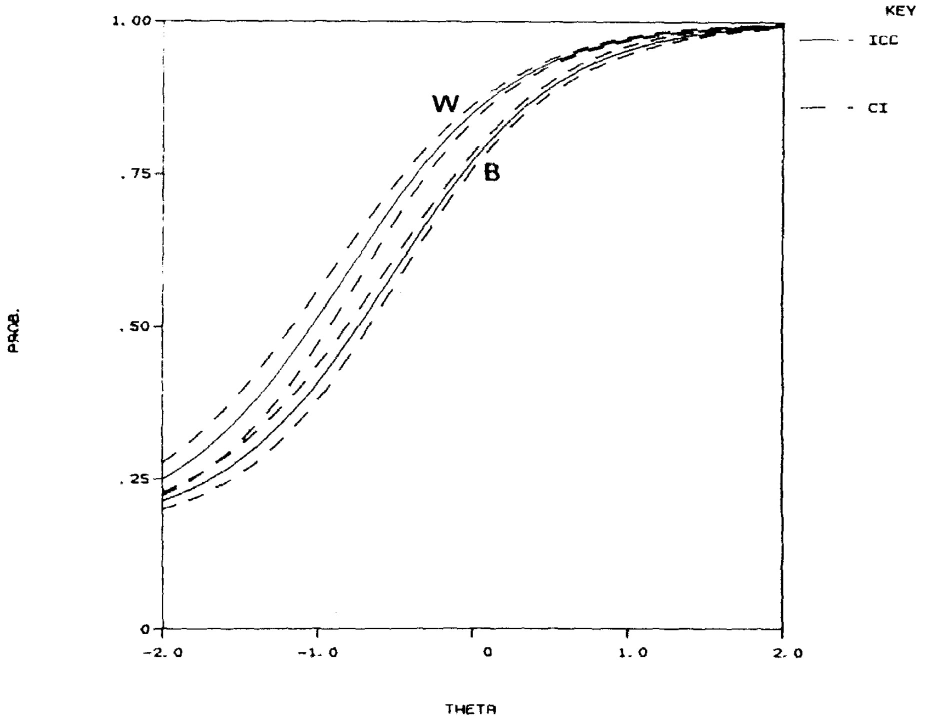


weighted in regions where more examinees are concentrated. In Figure 2a both the signed area and SOS4 index are large; whites have a considerable advantage over blacks for θ's above −1. In Figure 2b, the areas of advantage and disadvantage are more nearly equal, hence a near zero signed area. The SOS4 value for this same item is substantial, however, because more examinees of both groups, especially blacks, are located in the vicinity of −1 to 0 θ.

Item-characteristic curves for an item biased against *whites* are shown in Figure 3. Two graphs from the parallel analyses are presented for item 17 to illustrate the replication results. The amount of similarity between two independent but equivalent comparisons is fairly typical of the degree of stability found for consistently biased items (and for consistently unbiased items as well).

Item 30, in Figure 4, is an "artifactually biased" item. All of the indices are substantially deviant in comparison 1 but not in comparison 2. Item 30 is difficult for both groups. Hence, the *a* and *b* parameters must be estimated in a region where there is relatively little data. The difficulty in estimation is reflected in large standard errors. It should be noted, however, that even the

FIGURE 1b. Comparison of white and black item-characteristic curves for study 1, item 7 on the math test. (Example of an item found uniformly biased against blacks in both comparisons, i.e., ICCs do not cross.)

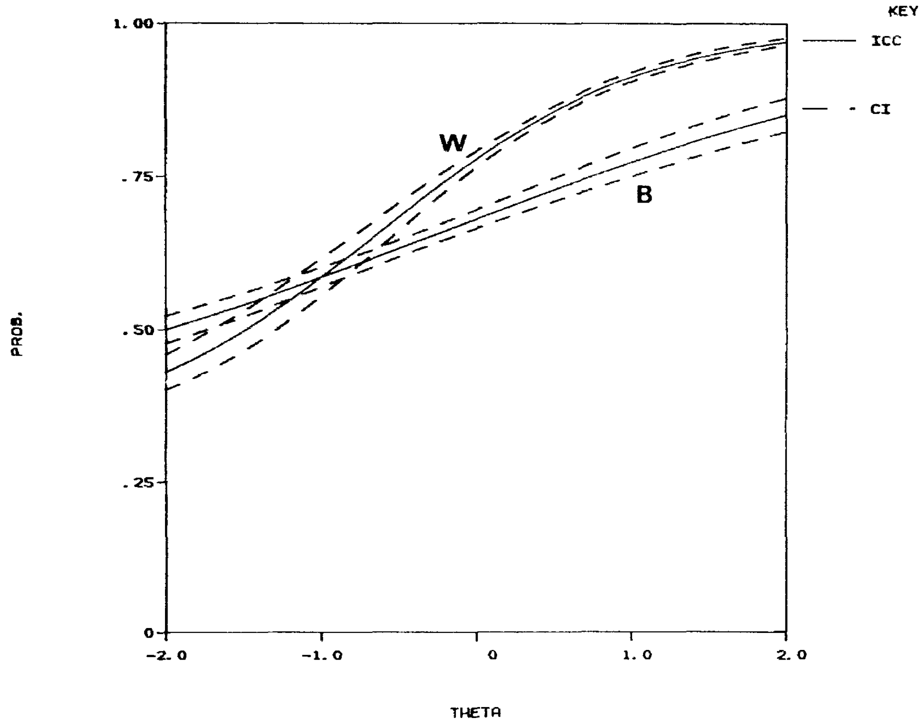


statistics that take standard errors into account (χ^2, SOS2, SOS4), and the SOS measures that deemphasize discrepancies in regions with little data, had large vaiues from this apparently spurious bias. Item 30 was a clear outlier, however, in the scattergram of *b*'s in the equating step for comparison 1; misestimation of *b*'s in *both* groups had a compound effect in comparison 1 that did not occur in comparison 2.

White-White and Black-Black Comparisons on the Math Test

Item-characteristic curves determined in two randomly equivalent groups should differ only by sampling error. Comparison 3 in Table I contains the bias results for two samples of whites (*W*1, *W*2). Logically, there should be no bias in this comparison and, indeed, inspection of these data indicates that all of the indices are appreciably smaller than in the white-black comparisons. Only item 30, which we know had estimation problems in sample 1, stands out with relatively large values. (Still, the numbers are much smaller than the corresponding indices in the between-ethnic comparison.)

FIGURE 2a. Comparison of white and black item-characteristic curves for study 2, item 6 on the math test. (Example of an item found to be predominantly biased against blacks in both comparisons, although the ICCs cross.)



The non-zero values of each index in comparison 3 indicate the ranges in magnitude that occur as sampling fluctuations. Therefore we used the largest value of each index occurring in study 3 as the cutoff for evaluating bias in the black-white studies.

The stability of results in the two white samples was relatively dramatic. Therefore, we wondered if the white-white comparison would produce too stringent a baseline. It was conceivable that estimation problems could be more difficult in the black group. Although all samples were equal in size ($n = 1{,}500$), something such as a range restriction problem in the black group could make parameters more unstable for this group. This unreliability could then lead to spuriously large bias indices—especially if the more stable white-white analysis were used as the baseline.

To test the above hypothesis we also conducted a black-black "bias" analysis. Indeed, we did encounter more estimation problems than with previous analyses. All but two item ICCs had been estimated for sample *B*1 when *c*'s were estimated in common with *W*1. However, standard errors could not be

FIGURE 2b. Comparison of white and black item-characteristic curves for study 1, item 11 on the math test. (Example of an item found to be biased against blacks in both comparisons. Although the ICCs cross, the item is biased against blacks in the region where most blacks are located.)

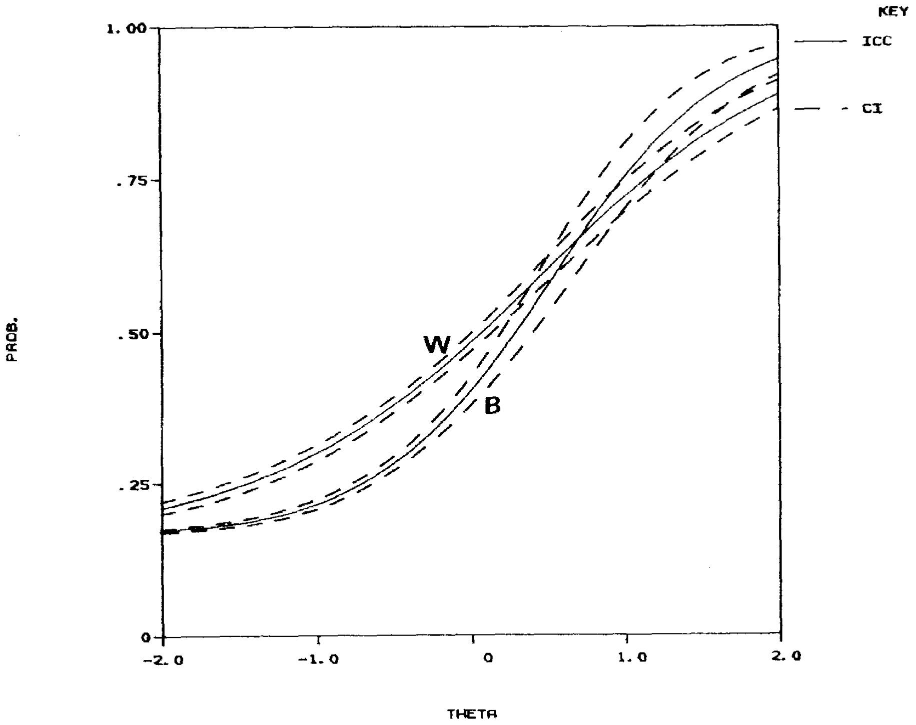


FIGURE 3. Comparison of white and black item-characteristic curves for item 17 on the math test for study 1 and study 2. (Example of an item found to be biased against whites in both comparisons, but by slightly different amounts.)

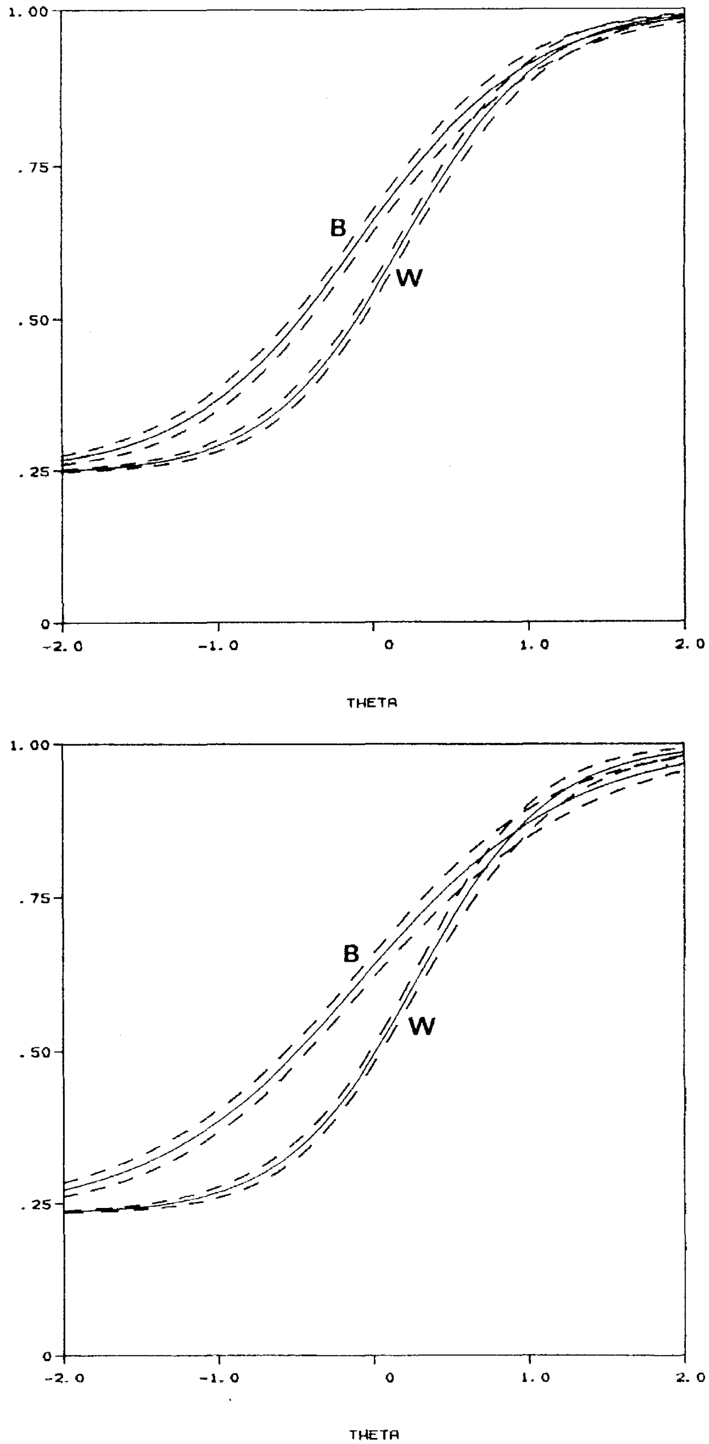
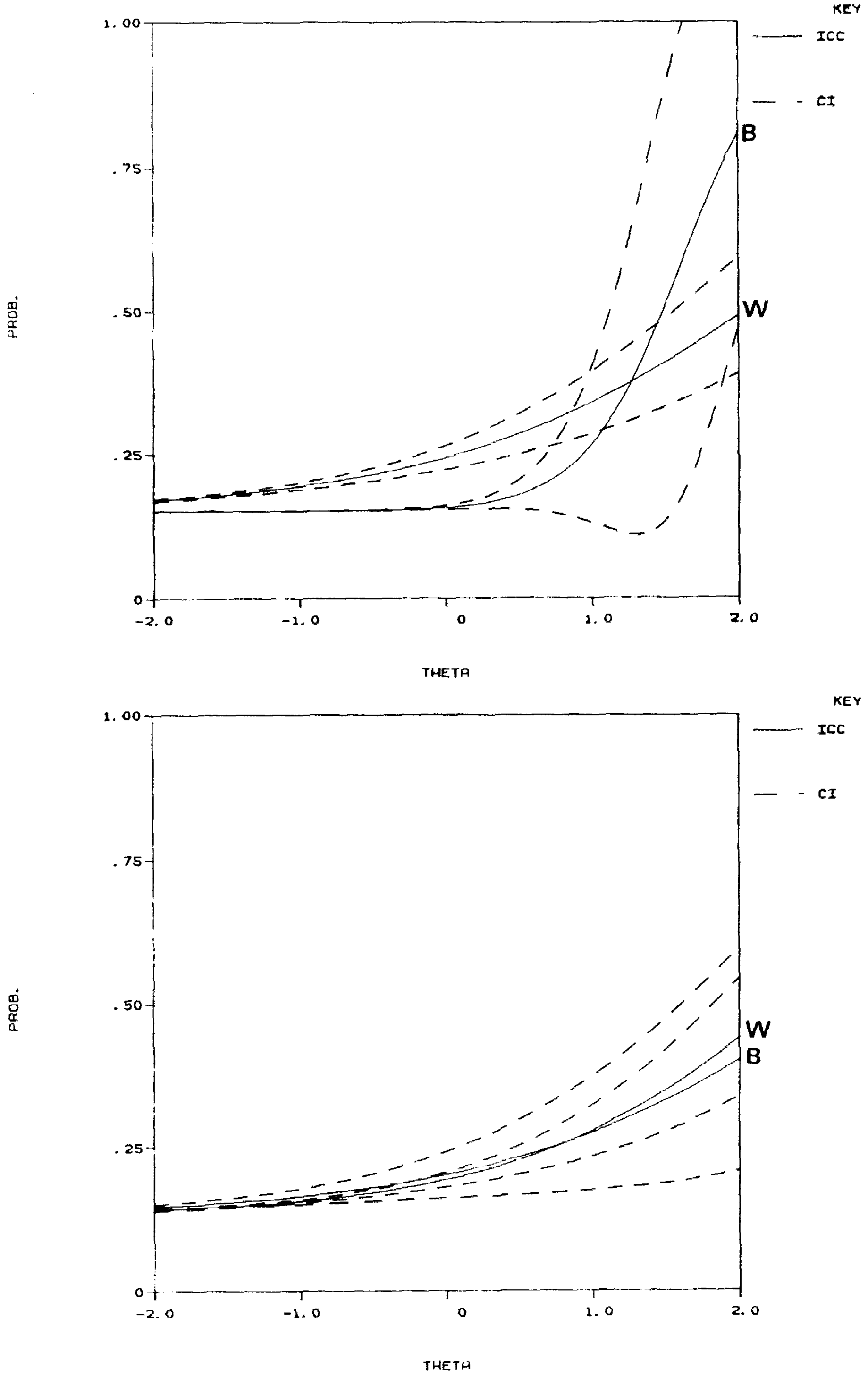


FIGURE 4. Comparison of white and black item-characteristic curves for item 30 on the math test for study 1 and study 2. (Example of an item found to be biased in comparison 1 but not in comparison 2.)



estimated for more than one third of the items when $B1$ was rerun with pooled c's from $B2$. Eventually we were able to finesse the LOGIST runs by inputting initial item parameters from the $B1$, $B2$ run and by raising the upper limits on a's. After these estimation difficulties were resolved, however, ICC comparisons for the two black samples (comparison 4, Table I) did *not* result in a wholesale increase in the number of large bias values. Therefore, we continued to use the baseline values obtained in the white-white study.

Pseudo-ethnic Comparison for the Math Test

It is conceivable that even IRT methods, which are theoretically sample invariant, may be inadequate when differences between groups are large. On the math total scores, blacks were $.91\sigma$ below the white group. To what extent might the apparent item discrepancies in Table I be due to failure of the model to cope with mean differences in the separate ICC analyses? To answer this question, we created a pseudo-black sample. This group was selected at random from the original file of white examinees but with the probability of being selected constrained to match the relative frequency distribution of black total math scores. (We recognize the circularity implicit in matching on the very test to be analyzed; in a separate program of research we are using different sets of background variables external to the test, e.g., SES factors and instructional history, to study their effect on the issue of bias.) The white sample matched to the black distribution can give us a rough idea of the amount of deviance showing up in the bias indices solely as a function of mean group differences and sampling error. Because of regression effects on individual items, however, the $W1$, Pseudo $B(W3)$ comparison is not quite so extreme as the $W1$, $B1$ difference.

The results of the pseudo-ethnic bias study are shown as comparison 5 in Table I. Note that there are very few large indices. Therefore, the large amount of deviance in the black-white analyses must be due to real differences in the functioning of the items across groups rather than to artifacts of the mean difference in math achievement.

The chi-square index produced the greatest number of large values in the pseudo-ethnic comparison. For the four items where the χ^2 is starred as biased but no other index exceeded its cutoff, there was in each case a fairly big shift in the b parameter. As Linn et al. (1980) point out, it is questionable whether differences only in b parameters should be taken as evidence of bias. For these items the b shift is not reflected in overall ICC differences, or else the other indices would have shown large effects as well.

Correlations and Agreements Among Bias Indices

In subsequent sections the bias analyses for the vocabulary test will be presented and the nature and importance of the apparent bias in the math test will be explored. Here, we wish to discuss some methodological issues regarding the functioning of the bias statistics. Results are presented for both tests to check on the generalizability of study findings.

To examine the relationships between indices, within-study correlations were obtained for each comparison on each test. Tables II and III contain the within-comparison coefficients for the math and vocabulary tests, respectively. As we explained in previous work (Shepard, Camilli, & Averill, 1981), Spearman rank-order correlations are preferred. With the Pearson *r*, one very extreme item will occasionally inflate or obscure the degree of relationship. When studying bias, congruence in the identification of extreme items is of primary interest; therefore, we did not wish to trim the distribution or eliminate outliers.

In Tables II and III, the first two entries are for comparisons where some bias is present. These are the between-ethnic comparisons. The remaining entries reflect the degree of correspondence between indices within a comparison where there are different amounts of sampling instability but presumably no bias. Although one might expect the correlations between indices to be higher in the presence of bias, this is not the case. The indices, which are similar to each other, are similar whether they are quantifying extreme deviance or only sampling perturbations. After all, whatever these sampling fluctuations are, they are constant *within a given comparison.*

The unsigned indices are highly correlated, suggesting they will yield fairly redundant information. The signed indices are also correlated with each other. However, the SOS4 statistic and the other signed statistics are less highly intercorrelated than the unsigned indices.

It was on the basis of these within-study correlations that we eliminated the simple sum-of-squares statistics from some of the results tables. The SOS1 index is essentially redundant with both the unsigned area and SOS2; SOS3 gives nearly the same picture of bias as the signed area. Note also that the pattern of relationships among indices (across comparisons of different types) was similar for both the math and vocabulary tests.

The more important test of the stability and validity of the indices as signs of bias is the pattern of correlations *between* study comparisons. The within-study correlations show consistency from both true and error sources of variance. The between-comparison correlations are given in Tables IV and V for the math and vocabulary test, respectively. Again, rank-order correlations were computed. These coefficients reflect how highly a bias index correlates with itself across study comparisons; that is, how consistently does it rank the 29 math items studied?

In a sense, these coefficients can be examined for convergent and discriminant validity as in a multitrait-multimethod matrix. The first line of each table is where we expect to see the effect of the trait on the magnitude of the

TABLE II

Intercorrelation[a] of Bias Indices Within Comparison on the Math Test (repeated for five comparisons)

Order of r's:		n's =
	B1, W1	30 items
	B2, W2	29
	W1, W2	27
	B1, B2	27
	W1, Pseudo B	29

	UA	SOS1	SOS2	X^2	SA	SOS3	SOS4	AD	BD
UA		.90 .89 .85 .71 .83	.83 .76 .78 .33 .73	.84 .81 .86 .65 .70	.26 .28 .26 .10 .55	.47 .36 .20 .02 .70	.47 .40 .07 .05 .55	-.40 .05 -.28 .22 -.27	-.22 -.02 .12 -.08 -.46
SOS1			.90 .84 .84 .53 .77	.94 .93 .98 .96 .89	.18 .23 .08 .06 .39	.37 .34 .11 .02 .59	.32 .34 .19 .15 .43	-.37 -.06 -.15 .00 -.23	-.17 -.03 .28 -.05 -.30
SOS2				.86 .92 .88 .57 .71	.06 .06 .24 .02 .18	.30 .19 .23 .00 .46	.32 .34 .21 .16 .39	-.44 -.18 -.32 -.02 -.38	-.03 .05 .13 -.02 -.11
X^2					.22 .14 .07 -.04 .32	.33 .23 .07 .02 .43	.18 .25 .17 .25 .25	-.44 -.16 -.16 -.10 -.18	-.24 -.03 .29 .06 -.26
SA						.84 .93 .86 .78 .86	.61 .70 .31 -.02 .58	-.02 .11 -.32 .20 .09	-.92 -.74 -.73 -.99 -.95
SOS3							.82 .82 .60 .41 .79	-.32 -.11 -.31 -.05 -.27	-.80 -.63 -.52 -.70 -.78
SOS4								-.25 -.09 -.17 -.56 -.30	-.47 -.29 .04 .08 -.46
AD									.11 -.02 .19 -.17 -.11

[a] All coefficients are Spearman rank-order correlations.

correlations. The trait is, of course, "bias" in the test items or differential functioning of the items due to cultural background. Only in the first row are the correlations between two randomly equivalent ethnic comparisons. Here we would expect to see consistency in the detection of bias. Indeed the degree of relationship is quite good; *r*'s for the math test range from .70 to .83; for the vocabulary test they are on the order of .60 to .86. (Note that *a* and *b* differences are presented to study their correspondence with other statistics, but they are not interpreted as indices of bias.)

The subsequent rows in the between-group matrices contain correlations where bias should *not* be the source of agreement. In all the remaining rows

TABLE III

Intercorrelation[a] *of Bias Indices Within Comparison on the Vocabulary Test (repeated for three comparisons)*

	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
UA		.87 .91 .95	.61 .78 .67	.71 .85 .95	.28 −.04 −.01	.39 .13 .11	.22 .37 .06	.42 −.14 −.16	−.31 .13 .01
SOS1			.81 .89 .76	.89 .97 .97	.15 −.07 .00	.26 .06 .16	.15 .25 .19	.48 .11 −.09	−.18 .07 −.02
SOS2				.94 .83 .80	−.07 .00 .22	.04 .16 .39	.07 .35 .40	.32 −.09 .08	.07 .08 −.23
χ^2					.06 −.15 .04	.14 −.03 .17	.14 .22 .17	.47 .21 −.08	−.04 .08 −.07
SA						.94 .89 .85	.53 .49 .12	.21 −.10 .25	−.96 −.90 −.97
SOS3							.72 .65 .36	.11 −.29 .13	−.92 −.81 −.82
SOS4								−.18 −.44 −.10	−.42 −.33 −.06
AD									−.18 −.12 −.20

Note. Order of *r*'s: *W*4, *B*4; *W*5, *B*5; *W*4, *W*5.
n's = 25 items; 21 items; 21 items.

[a] All coefficients are Spearman rank-order correlations.

at least one or both of the comparisons were between equivalent groups (either both white or both black). These correlations should show discriminant validity or the lack of method-specific correlations. These correlations should be near zero, confirming a lack of bias when none exists conceptually. However, it should be noted that these pairs of comparisons do share some consistent errors because one sample is repeated in both comparisons. For example, we expect the correlation between indices obtained in the *W*1, *B*1 study and those from the *B*1, *B*2 study to correlate zero. Bias can be present in the first

TABLE IV
Correlations[a] *of Each Bias Index with Itself Across Study Comparisons on the Math Test*

	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
*W*1, *B*1 with *W*2, *B*2	.71	.71	.72	.80	.72	.80	.75	.83	.65
*W*1, *B*1 with *W*1, *W*2	.33	.14	.16	−.02	.08	.28	.25	−.20	.22
*W*2, *B*2 with *W*1, *W*2	.27	.12	.03	.08	.01	−.04	−.12	−.46	.08
*W*1, *B*1 with *B*1, *B*2	.32	.00	.26	.17	−.11	−.03	−.13	−.21	−.15
*W*1, *B*1 with *W*1, Pseudo *B*	.32	.26	.33	.33	.49	.26	.37	.28	.41
*W*1, *W*2 with *W*1, Pseudo *B*	.24	.22	−.13	.19	.17	.39	.32	.16	.21

Note. Only for the correlations between *W*1, *B*1 with *W*2, *B*2 is there the possibility for agreement when bias is present. For other correlations, one or both of the comparisons involved randomly equivalent groups or two white groups; therefore, there should be no consistent bias. These latter pairs do share some consistent errors, however, because in each case one of the samples is repeated in both comparisons. Only in the correlations below should there be both no bias and no sample redundancy.

	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
*W*1, *W*2 with *B*1, *B*2	.32	.22	−.03	.21	.42	.02	−.15	−.04	.09

[a] All coefficients are Spearman rank-order correlations.

TABLE V
Correlations[a] *of Each Bias Index with Itself Across Study Comparisons on the Vocabulary Test*

	UA	SOS1	SOS2	χ^2	SA	SOS3	SOS4	AD	BD
*W*4, *B*4 with *W*5, *B*5	.60	.69	.85	.83	.63	.84	.78	.32	.56
*W*4, *B*4 with *W*4, *W*5	.60	.53	.18	.45	.00	−.03	.12	.32	−.14
*W*5, *B*5 with *W*4, *W*5	.61	.47	.24	.45	−.49	−.31	−.08	−.33	−.50

Note. Only for the correlations between *W*4, *B*4 with *W*5, *B*5 is there the possibility for agreement when bias is present. The latter pairs do share some consistent errors, however, because in both cases one of the white samples is repeated in both comparisons.

[a] All coefficients are Spearman rank-order correlations.

study but not the second. The two comparisons do, however, share the *B*1 sample. Therefore, the two studies could have some consistent spurious "bias" based on sample characteristics. Only in the last row of the math data (Table IV) are there correlations between conditions where there should be both no bias *and* no consistent sampling error.

The discriminant coefficients show the reduced relationships necessary to support the validity of the bias indices. For example, on the math test the SOS2 statistic is correlated .72 with itself when bias is present in both studies; it is correlated only .03 to .33 across studies where bias is not the source of relationship. However, the pattern of high-trait, low-method correlations is not so good for the unsigned indices on the vocabulary test. Two reasons should be kept in mind: The vocabulary test is less biased, and as we explained in previous research (Shepard, Camilli, & Averill, 1981), it is more difficult to show ranking consistency with unsigned indices because they are one-tailed distributions. That is, unsigned indices have both items biased against blacks and whites in the same tail of the distribution, making it more difficult to demonstrate consistency across parallel studies.

We are tentatively prepared to recommend the SOS2, SOS3, and SOS4 indices as the more valid indices of bias. Not only are these statistics the most consistent in detecting bias in the ethnic comparisons, but they also intercorrelate the least in situations of no bias. A minor caveat is warranted, however, regarding the two weighted measures (SOS2 and SOS4). Because in our method of IRT estimation we fixed *c*'s from a common analysis, we assumed that standard errors for *c* were zero in the weighted SOS formulae. To the extent that this assumption was erroneous, especially for very easy items, the same false assumption could add spurious agreement to the between-study consistency. We judged this effect to be very slight. This problem could not, of course, explain the desirable drop-off in correlations in contrasts where bias should not be present. The SOS3 statistic was not affected by this assumption.

Correlation coefficients are only a crude method for summarizing the consistency of indices in identifying biased items. We are not interested in the consistency with which unbiased items are ranked. Rather, only the consistency at the extremes of the item rankings is important. Using the cutoffs determined from the white-white analysis, items were classified as either biased or not biased by each index. The contingency tables in Table VI show the consistency of these dichotomous classifications from one black-white comparison to the other (on the math test). Here it should be clear that the SOS2 and SOS4 are relatively the best, and in an absolute sense, quite good at consistently classifying items as biased or not biased. The χ^2 statistic is next-best in the amount of replicated bias. But, as we explained earlier, the χ^2 can consistently identify as biased items that have a large parameter difference but do not have a commensurate probability difference for most sampled θ's. This occurs especially when items have large b differences at the extreme ranges of θ. The χ^2 index has the property of consistency but is less desirable on other grounds.

TABLE VI

Agreement of Indices in Equivalent White-Black Comparisons on the Math Test

Unsigned Area:

		Comparison 1	
		B	NB
Comparison 2	B	9	2
	NB	3	15

83% agreement

χ^2:

		Comparison 1	
		B	NB
Comparison 2	B	13	2
	NB	2	12

86% agreement

SOS1:

		Comparison 1	
		B	NB
Comparison 2	B	12	2
	NB	2	13

86% agreement

SOS2:

		Comparison 1	
		B	NB
Comparison 2	B	8	1
	NB	2	18

90% agreement

Signed Area:

		Comparison 1	
		B	NB
Comparison 2	B	7	4
	NB	1	17

83% agreement

SOS3:

		Comparison 1	
		B	NB
Comparison 2	B	11	3
	NB	2	13

83% agreement

SOS4:

		Comparison 1	
		B	NB
Comparison 2	B	7	1
	NB	2	19

90% agreement

Note: These counts are based on the individual item data presented in Table 1. Biased items, starred in Table 1, had indices for a given comparison that exceeded the cut-off value determined from the white-white comparison. For the X^2 index, however, the critical value of 5.99 was used here.

The agreement results found for the math test were only partially duplicated on the vocabulary test. The percentages of agreements were as follows: UA, 70%; SOS1, 75%; SOS2, 85%; χ^2, 85%; SA, 70%; SOS3, 75%; and SOS4, 85%. On the vocabulary test there was less bias; also on this test we had more difficulty justifying a particular cutoff from the white-white analysis.

Substantive Interpretation of Bias Results for the Math Test

The original premise motivating this research was that the results of item bias analyses would be more interpretable if statistical artifacts could be controlled. Specifically, we expected to see more of a pattern in test items found to be biased if we studied only those items that were cross-validated, if we found them to be deviant in parallel black-white comparisons.

Once we had identified the consistently biased and unbiased items on the math test, we looked at the actual test questions. It was immediately apparent that verbal math problems were the source of the bias against blacks. (The cross-validation did little to clarify this picture. The indices were consistent enough across studies that similar insight would have been gained by looking only at the results from one black-white comparison.)

All of the HSB math items (Part 1 and Part 2) had the following format:

Directions: Each problem in this section consists of two quantities, one placed in Column A and one in Column B. You are to compare the two quantities and mark oval

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the size of the relationship cannot be determined from the information given.

	Sample Questions		Sample Answers			
	Column A	Column B				
Example 1.	20 percent of 10	10 percent of 20	A	B	C	D
Example 2.	6 × 6	12 + 12	A	B	C	D

Answer C is marked in Example 1 because the quantity in Column A is equal to the quantity in Column B. Answer A is marked for Example 2 because the quantity in Column A is greater than the quantity in Column B.

We called items similar to Example 1 verbal and those similar to Example 2 numeric.

A more realistic illustration of the verbal-type items is provided by the two following questions. These items were written to parallel two actual test questions that were found to be consistently biased against blacks:

	Column A	Column B
1.	Number of centimeters between −7 cm and + 8 cm	Number of centimeters between −8 cm and + 7 cm
2.	Cost per pound at a rate of $4.00 for twenty pounds	Cost per pound at a rate of 3 pounds for 60¢

A type of numeric problem found to be consistently biased in favor of blacks was parallel to the following example:

3.	326	$3(10)^3 + 2(10)^2 + 6(10)$

Numeric items that were consistently unbiased were similar to the following:

4.	$\sqrt{16}$	16
5.	5a	6x

The only numeric item found to be biased against blacks was comparable to this item:

6.	$33 \div 5$	$37 \div 5$

If questions had a verbal phrase in one column and a numeral in the other column, we called them $V + N$. The classification of math items as verbal or numeric was shown in Table I.

Table VII is a contingency table depicting the cross-tabulation of the bias results with item type. These data show a striking degree of relationship suggesting that the bias indices are indeed sensitive to a change in meaning of the underlying trait for black examinees as measured by the verbal items.

The foregoing conclusions have been rather enthusiastic. The bias indices, especially the SOS statistics, yield consistent results (with these sample sizes). They show appreciable discriminant validity between the biased and nonbiased studies. And, when the test questions themselves are examined, the indices seem to have signaled interpretable instances of differential performance. This enthusiasm must be tempered somewhat by the following result.

TABLE VII
Bias Classification and Item Type For Math Items

	Verbal	$V + N$	Numeric
Consistently biased against blacks	6	0	1
Not biased	5	2	12
Consistently biased against whites	0	0	3

In practical terms we wished to quantify the effect of having biased items in the test. Therefore, we rescored the math test, deleting the seven items found to be consistently biased against blacks. We compared the new black and white means in the metric of the white standard deviation. The difference was .81σ. For the unexpurgated test it had been .91σ. The effect of the biased items (however consistent) is small but not trivial.[4] The relatively small magnitude of the bias effect can also be seen by examining the ICC graphs for typical biased items. Although the curves are discernably different, the probability differences are not very large. Item 6, comparison 2, was selected for illustration (Figure 2a) because it had the largest unsigned area statistic of all the biased items. At its height the probability difference between blacks and whites is .13. More typically the largest black-white difference on a biased item is only .05 to .10. This would mean on average roughly one more item correct for blacks if the biased items were removed.

To illustrate further the practical import of the seven items biased against blacks, we also simulated the effect on failure rates if the test had been used to make pass/fail decisions as in a minimum-competency testing program. To establish comparable cutoff scores, raw scores were selected that would fail 10% of the whites on both the full and debiased tests. The corresponding failure rates for blacks on the two tests were 36.3% and 30.3%, respectively.

The finding that the overall effect of bias is small tempers both our methodological and substantive conclusions. We must remember that internal bias indices cannot detect constant bias. Because the format of all the HSB problems requires some verbal reasoning, we may have underestimated the effect of pervasive bias from this source. It is also plausible that a math achievement instrument developed for a national survey would be much less biased than many other tests. Because the bias results were consistent and interpretable in a test with a relatively small bias effect, we are inclined to believe that the indices are sensitive to relatively subtle but meaningful sources of differential performance. We expect that the desirable properties of the indices for bias detection would be enhanced in situations where there was a greater amount of bias. We would predict, for example, that in field trials of new test items there would be more bias to be detected.

Bias Results for the Vocabulary Test

Bias indices for the vocabulary test are presented in Table VIII. Again, comparisons 1 and 2 are randomly equivalent black-white analyses. Comparison 3 is between two random samples of whites, a circumstance where there

[4] The effect on black-white differences would have been smaller still if we had deleted the three items biased against whites. However, the bias against whites was far less interpretable.

TABLE VIII

Signed and Unsigned Bias Indices for Vocabulary Items in Three Comparison Studies

	Comparison 1: W4, B4					Comparison 2: W5, B5					Comparison 3: W4, W5				
	Unsigned			Signed		Unsigned			Signed		Unsigned			Signed	
Item	UA	SOS2	X^2	SA	SOS4	UA	SOS2	X^2	SA	SOS4	UA	SOS2	X^2	SA	SOS4
1															
2	.26*	11.04*	23.31*	-.26*	-11.04*	.19	10.74*	15.00*	-.18	-10.73*	.05	.23	1.56	-.05	-.15
3	.02	.43	.43	.02	.43	.06	.78	1.70	.06	.78	.04	1.54	.96	.01	1.53
4	.29*	8.79*	17.66*	.15	6.82*	.32*	12.08*	23.22*	.32*	12.08*	.19	2.95	7.67*	-.13	-.06
5	.07	4.70	4.24	-.06	-4.67	.07	.90	2.12	-.07	.90	.05	1.04	.58	.02	-1.02
6	.29*	8.85*	16.09*	-.21	-8.71*	.33*	10.50*	23.05*	-.33*	-10.50*	.14	2.48	4.99	.13	2.48
7	.02	.04	.06	.00	.01	.07	.50	1.61	.07	.50	.03	.24	.38	.02	.21
8	.42*	3.41	8.89*	.42*	3.41	.29*	26.75*	7.19*	.01	26.68*	.24*	7.24*	7.19*	.14	-6.74*
9	.10	1.89	2.72	.10	1.89	.29*	4.87	9.98*	.16	-3.03	.14	2.18	4.78	-.04	.73
10	.19	2.19	5.12	-.11	-1.06	.21	3.53	6.14	-.06	1.70	.03	.07	.21	.02	.07
11	.15	7.91*	7.49*	-.11	-7.89*	.15	7.14	5.31	-.06	-7.12*	.03	.15	.29	.03	-.12
12	.31*	16.50*	10.28*	-.08	15.65*						.25*	19.89*	11.79*	.00	19.28*

13	.28*	3.40	5.22	-.25*	-3.40										
14	.18	1.98	3.21	-.18	-1.98	.14	2.61	2.98	-.13	-2.61	.10	.16	1.22	-.09	-.15
15	.21	3.20	6.45	.21	3.20	.18	7.21	5.99	.04	7.06*	.12	1.12	3.76	.10	1.06
16	.20	5.94	14.46*	-.07	5.72	.39*	5.38	34.99*	-.30*	4.02	.24*	.80	5.99	.23	-.44
17	.73*	25.65*	39.76*	.73*	25.65*	.69*	29.20*	53.31*	.67*	29.18*	.90	.42	.58	-.01	.38
18	.24*	18.67*	35.29*	.02	9.78*	.17	7.18	13.94*	-.01	2.01					
19	.23	1.78	3.95	.08	.72	.17	2.14	3.79	.16	2.13	.05	.22	.53	-.05	-.22
20	.08	.79	2.37	-.08	-.79										
21						.17	6.67	.77	.17	6.67	.24*	.15	2.65	-.24*	-.08
22	.27*	1.03	4.96	.27*	.92	.03	.04	.21	.03	.03	.10	1.10	2.10	-.04	.74
23	.03	.13	.31	-.03	-.13	.08	.62	1.39	.06	.61	.09	.46	1.09	-.03	.00
24	.34*	1.63	3.75	.22	-1.07	.20	1.70	1.88	-.12	1.54	.13	3.02	3.58	.13	3.02
25	.14	1.39	.90	-.06	1.35	.04	.01	.08	.04	.01	.02	.11	.13	.02	.11
26	.22	6.88	10.82*	.09	-6.65										
27	.22	6.18	8.55*	.09	-5.97										

Note: To establish a baseline for judging the magnitude of the bias indices, the values from the second most deviant item in the white-white comparison were used. Indices that exceeded this cut-off in other comparisons are starred as "biased." For the sake of consistency, the item 8, W1, W2, χ^2 of 7.19 was used; however, 5.99 is the critical value for statistical significance at α = .05.

should be no bias. The largest values obtained in the white-white comparison were used as baselines for interpreting the size of indices in the between-ethnic comparisons. Because two items in the white-white analysis stood out as different from the typical range of values, the indices from the second-most discrepant item were used to establish the cutoffs.

The methodological results from the vocabulary test were discussed earlier. Generally, they corroborated the findings based on the math test, but patterns were sometimes weaker because there was overall less internal bias in the vocabulary test. This test was difficult for both groups. Inspection of the content also suggested that the test was extremely unidimensional; for example, we could not categorize the words a priori as being more or less frequent in everyday language. All of the words seemed to have a literary flavor and were school and book oriented. Therefore, we were uncertain as to whether the analysis would detect differential difficulty.

The consistently biased items seen in Table VIII are not immediately interpretable. Initially we conjectured that there might be some speed effects present in this test because the two parts had time limits of only 5 and 4 minutes, respectively. (Note that Part II starts with item 16.) However, there were not, in fact, appreciably different omitted or not-reached rates between the two groups. Four items appear to be consistently biased against blacks: items 4, 16, 17, and 18. This result was puzzling because these are consistently the easiest items in the test. Only three other items (items 1, 3, and 5) are as easy (and item 1 could not be estimated). Apparently there may be a floor effect here whereby blacks scoring near chance on many other items in the test cannot look as different on the very difficult items as they do on easy items. (Note, item 8 would have contradicted this trend because it is biased against blacks and is difficult [$P_W = .35$; $P_B = .20$]; however, we ignored item 8 because it was also biased in the white-white comparison.)

Summary and Conclusions

The purpose of this research was to apply item response theory bias detection procedures to both a mathematics achievement and vocabulary test. Because the results of previous item-bias studies often have been uninterpretable, we wished to account for statistical artifacts by conducting cross-validation or replication studies. Therefore, each analysis was repeated on randomly equivalent samples of blacks and whites. Furthermore, to establish a baseline for judging bias indices that might be attributable only to sampling fluctuations, bias analyses were conducted comparing randomly selected groups of whites. Also, to assess the effect of mean group differences on the appearance of bias, pseudo-ethnic groups were created. That is, samples of whites selected to simulate the average black-white difference were also tested for bias.

The validity and sensitivity of the IRT bias indices were supported by several findings:

1. A relatively large number of items (10 of 29) on the math test was found to be consistently biased; the results were replicated in parallel analyses. (Seven were biased against blacks, three were biased against whites.)
2. The bias indices were substantially smaller in white-white analyses. That is, with the exception of one or two estimation artifacts, indices did not find bias in situations of no bias.
3. Furthermore, the indices (with the possible exception of χ^2) did not find bias in the pseudo-ethnic comparison. Therefore, bias by these methods is not an artifact of mean-group differences.
4. The pattern of between-study correlations showed high consistency between analyses where bias was plausibly present (i.e., between parallel ethnic comparisons).
5. Also, the indices met the discriminant validity test. That is, the correlations were low between conditions where bias should not be present.
6. For the math test, where a substantial number of items appeared biased, the results were interpretable. Verbal math problems were systematically biased against blacks.
7. The desirable pattern of between-comparison correlations was replicated on the vocabulary test, albeit somewhat weaker because of less bias on this measure.

Overall, the sums-of-squares statistics (weighted by the inverse of the variance errors) were judged to be the best indices for quantifying ICC differences between groups. Not only were these statistics the most consistent in detecting bias in the ethnic comparisons, but they also intercorrelated the least in situations of no bias. Lord's (1980) asymptotic chi-square was consistent but was sometimes sensitive to parameter differences that did not have corresponding effects on ICC differences.

When statistically biased items on the math test were examined, a strong relationship was found between the verbal properties of the item and bias classification. Most of the verbal problems on the test were biased against blacks, and with one exception, numeric problems were not. This highly reliable and interpretable result had to be tempered by the finding that the magnitude of the bias effect was relatively small. When items biased against blacks were deleted and the test rescored, the difference between blacks and whites was changed from $.91\sigma$ to $.81\sigma$. The bias indices are apparently sensitive to consistent but subtle effects. Presumably the validity evidence for the bias statistics would be increased in situations where there is greater bias, as in field tests of newly developed test items. We did not make substantive interpretations of bias findings for the vocabulary test. The amount of internal bias was far less for this instrument.

Acknowledgments

We wish to thank the Council on Research and Creative Work and Dean Richard Turner, School of Education, University of Colorado, for their financial support. We are grateful to Dr. Samuel Peng and Jeffrey Owings for providing a pupil file of individual item data as well as the publicly available aggregate tapes for pupils, parents, and schools.

We are grateful to Dr. Thomas Hilton and to the Educational Testing Service for providing us with copies of the HSB test materials. We thank the Educational Testing Service and National Commission on Educational Statistics for permission to reproduce the sample items and to create parallel item types to illustrate the nature of the verbal and numeric problems.

References

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (Ed.). (1982). *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Bond, L. (1981). Bias in mental tests. In B. F. Green (Ed.), *Issues in Testing: Coaching, Disclosure and Ethnic Bias.* San Francisco: Jossey Bass.

Bougon, M. G., & Lissak, R. I. (1981, August). *Detecting item bias using the three-parameter logistic model and the B and C statistics.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book.* Boston: Houghton Mifflin.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36,* 1067–1077.

Divgi, D. R. (1981, April). *Potential pitfalls in applications of item response theory.* Paper accompanying discussant comments at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Green, D. R., Coffman, W. E., Lenke, J. M., Raju, N. S., Handrick, F. A., Loyd, B. H., Carlton, S. T., & Marco, G. L. (1982). Methods used by test publishers to debias standardized tests. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Green, D. R., & Draper, J. F. (1972, September). *Exploratory studies of bias in achievement tests.* Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Harman, H. H. (1967). *Modern Factor Analysis.* Chicago: The University of Chicago Press.

Hunter, J. E. (1975, December). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Ironson, G. H., & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209–225.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 90,* 185–244.

Jensen, A. R. (1976). Test bias and construct validity. *Phi Delta Kappan, 58,* 340–346.

Jensen, A. R. (1977). An examination of cultural bias in the Wonderlic Personnel Test. *Intelligence, 1,* 51–64.

Linn, R. L. (1982, August). *Selection bias: Multiple meanings.* Paper based on a presidential address to the Division of Evaluation and Measurement at the annual meeting of the American Psychological Association, Montreal.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension* (Tech. Rep. No. 163). Urbana, IL: University of Champaign-Urbana, Center for the Study of Reading. (ERIC Document Reproduction Service No. ED 184 091).

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic Problems in Cross-Cultural Psychology.* Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Merz, W. R., & Grossen, N. E. (1979). *An empirical investigation of six methods for examining test item bias.* (NIE Grant No. 6-78-0067). Sacramento: California State University.

Petersen, N. S. (1977, June). *Bias in the selection rule: Bias in the test.* Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40,* 397–404.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213–233.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1–10.

Sandoval, J., & Miille, W. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48,* 249–253.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Shepard, L. A. (1981). Identifying bias in test items. In B. F. Green (Ed.), *Issues in Testing: Coaching, Disclosure and Ethnic Bias.* San Francisco: Jossey Bass.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias.* Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317–375.

Wood, R. L., & Lord, F. M. (1976). *A User's Guide to LOGIST.* Research Memorandum. Princeton, NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters.* Research Memorandum. Princeton, NJ: Educational Testing Service.

Authors

LORRIE SHEPARD, Associate Professor, Laboratory of Educational Research, Box 249, University of Colorado, Boulder, CO 80309. *Specializations:* Measurement and evaluation.

GREGORY CAMILLI, Research Associate, Human Systems Institute, P.O. Box 1761, Boulder, CO 80306. *Specializations:* Program evaluation and applied statistics.

DAVID M. WILLIAMS, Graduate Student, Department of Psychology, Box 345, University of Colorado, Boulder, CO 80309. *Specializations:* Cognitive and quantitative psychology.