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

IN ITEM BIAS RESEARCH

LORRIE SHEPARD
University of Colorado

GREGORY CAMILLI
Human Systems Institute

and

DAVID M. WILLIAMS
University of Colorado

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 sub-

will produce invalid indices of bias in the presence of group mean differences

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-

Merz & Grossen, 1979; Rudner, Getson, & Knight, 1980b). Because the

without replacement, so the samples were independent.)

Comparison 3: $W1$, $W2$ white samples from comparison 1 and compari-

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

Scale Equating

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 differ[illegible]

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.

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.

$$SOS3_i = \frac{1}{n_W + n_B} \sum_{j=1}^{n_W + n_B} \{\hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j)\} \left| \hat{P}_{iW}(\theta_j) - \hat{P}_{iB}(\theta_j) \right|,$$

where terms are as defined previously.

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

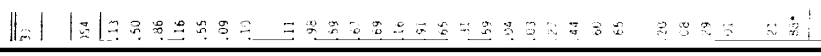
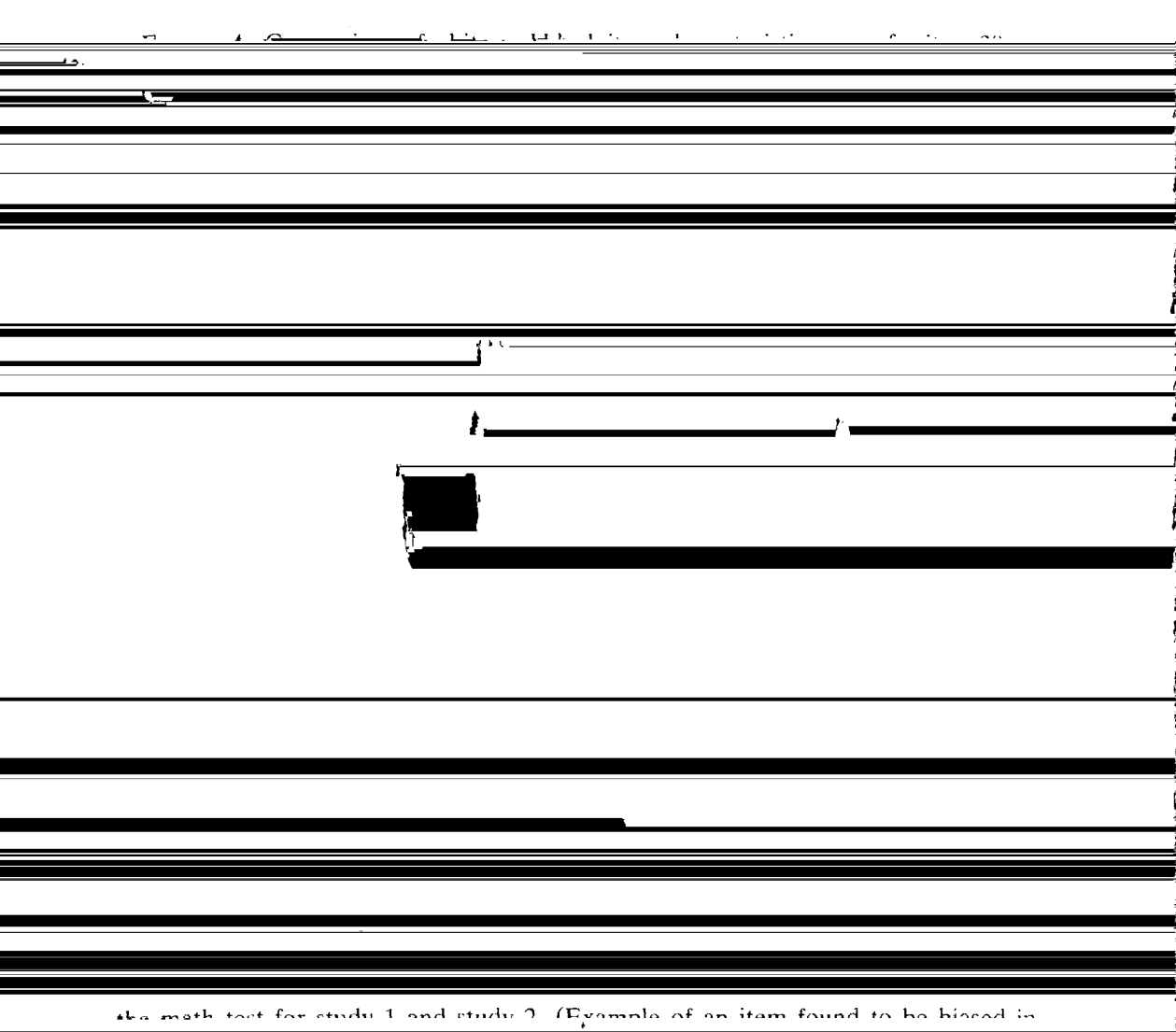


Figure 1. 4-pra item characteristic curves for blacks and whites on several

weighted in regions where more examinees are concentrated. In Figure 2a both the signed area and SOS4 index are large; whites have a considerable

The mean zero values of each index in comparison 3 indicate the ranges in

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 in

estimated for more than one third of the items when $B1$ was rerun with pooled

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 respec-

TABLE II

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

functioning of the items due to cultural background. Only in the first row are

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 $W1$, $B1$ study and those from the $B1$, $B2$ study to correlate zero. Bias can be present in the first

TABLE IV

Correlations[a] *of Each Bias Index with Itself Across Study Comparisons*

ne The agreement results found for the math test were only partially duplicated

	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¢

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

Studies

omparison 3: W4, W5				
ed		Signed		
OS2	χ^2	SA	SOS4	
.23	1.56	-.05	-.15	
.54	.96	.01	1.53	
.95	7.67*	-.13	-.06	
.04	.58	.02	-1.02	
.48	4.99	.13	2.48	
.24	.38	.02	.21	
.24*	7.19*	.14	-6.74*	
.18	4.78	-.04	.73	
.07	.21	.02	.07	
.15	.29	.03	-.12	
.89*	11.79*	.00	19.28*	

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

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.

Acknowledgments

We wish to thank the Council on Research and Creative Work and Dean Richard

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 Mo*

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.