The Reliability of Six Item Bias Indices

H. D. Hoover
The University of Iowa

Michael J. Kolen
American College Testing Program

The reliabilities of six item bias indices were investigated for each of the eleven tests of the Iowa Tests of Basic Skills, using random samples of fifth-grade students. The reliability of an index was defined as its stability from one randomly equivalent group to another. Both racial and sexual bias were considered. In addition, correlations among bias indices were investigated. The results indicate that the item bias indices investigated were fairly unreliable when based on sample sizes of 200 minority and 200 majority examinees. Consequently, this study suggests that, with sample sizes of about 200, the use of item bias indices to screen achievement test items cannot be expected to lead to consistent decisions about which items are biased.

Biased items can be eliminated from achievement tests in two stages. First, "experts" can judge the fairness to various groups of the presentation format and content of potential test items. Items judged to be unfair, or biased, can then be excluded from final test forms. Then item bias indices can be used to screen potential test items as a second stage. Ideally, bias indices are calculated from item tryout data. Items identified as biased by the indices can then be excluded from final test forms in much the same way that items with low item discrimination indices are usually eliminated during the item tryout stages of test development.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 2, Spring 1984, pp. 173–181

0146-6216/84/020173-09$1.70

If they are to be useful for screening purposes, item bias indices should produce stable results. However, research has suggested that item bias statistics may be fairly unstable. Scheuneman (1980) and Linn, Levine, Hastings and Wardrop (1981) found only modest agreement among item bias indices across independent samples. Linn et al. (1981) speculated that "it may be difficult to identify biased items because of the unreliability of the indices used" (p. 170). From a literature review, Ironson (1982) concludes that "we need more information on the reliability of the [item bias detection] methods" (p. 152). Studies by Plake (1980) and Qualls and Hoover (1981) found that statistical measures of item bias are essentially uncorrelated with item bias judgments made by experts. These research studies, however, do not directly address the issue of the reliability (stability from one randomly equivalent group to another) of item bias indices.

The present paper focuses on whether item bias indices are sufficiently stable to be useful in the identification of potentially biased items in the pretest phase of test development. A realistic situation was designed which included:

1. Sample sizes which were about as large as those typically available for minority groups in pretest situations;
2. Selecting examinees in such a way that the influence of curricular or instructional bias (Linn & Harnisch, 1981, p. 116) was minimized;
3. Studying a number of tests to allow for some generalizability of results; and

4. Studying tests in which item bias statistics were *not* used in their development.

This study investigated the reliabilities of each of six internal-criterion item bias indices, using both race and sex categorizations, for each of the eleven tests of the Iowa Tests of Basic Skills (ITBS). Only unsigned versions of the indices (Ironson & Subkoviak, 1979) were investigated since item screening, as usually conceived, involves eliminating items biased against any group.

No discussion of the differences among definitions of item bias or among item bias statistics will be presented here. These issues are discussed in a variety of sources including Berk (1982), Hunter (1975), Ironson and Subkoviak (1979), Lord (1980), Rudner, Getson, and Knight (1980a, 1980b), Marascuilo and Slaughter (1981), and Shepard, Camilli, and Averill (1981).

Item Bias Indices

Six different item bias indices were investigated in this study. The *Difficulty* and *Delta* indices are designed to detect group differences (e.g., between blacks and whites) in relative item difficulty. The *Biserial* and *Point Biserial* indices are designed to detect group differences in item discrimination. The *Scheuneman* and *3-Parameter* indices are designed to detect differences in relative item difficulty by score level and latent ability level, respectively.

Difficulty and Delta Indices

The Difficulty index used in this study is the absolute value of the transformed item difficulties (45° line method) described by Rudner et al. (1980a). For this index, the within-group item difficulties are standardized (mean of zero; standard deviation of one). The Difficulty index for an item is the absolute value of the difference between the standardized item difficulties for the two groups. The Delta index for an item is similar to the Difficulty index except that the within-group item difficulties are first transformed using an inverse normal transformation. A similar approach was used by Angoff and Ford (1973).

Biserial and Point Biserial Indices

The Biserial index for an item is the absolute difference between the within-group biserial correlations of the item with total score. The Point Biserial index for an item is the absolute difference between the within-group point biserial correlations of the item with total score.

Scheuneman Index

The Scheuneman index (Scheuneman, 1979) was calculated for each item using five score levels. The score levels were defined such that approximately equal numbers of examinees were in each level. Scheuneman (1979) claims that her index could be expected to be distributed approximately as chi-square with four degrees of freedom when five score levels are used.

3-Parameter Index

The 3-Parameter index is a modification of the index proposed by Linn and Harnisch (1981). This index was chosen because it can be used with smaller sample sizes than the more widely recommended item characteristic curve index suggested by Lord (1980). For the 3-Parameter index, first the item and ability parameters of the three-parameter logistic model are estimated for the combined group of examinees. For example, item responses for black and white students are pooled in order to estimate the model parameters. The two groups of examinees are then separated. For each examinee, the difference between the examinee's estimated probability (p) of correctly answering the item and the examinee's actual response to the item (1 = correct; 0 = incorrect) is found. This quantity is then divided by $[p(1 - p)]^{1/2}$ and averaged over examinees within each group. The mean for each group is then squared and the two squared means summed to arrive at the 3-Parameter index.

Method

The data consisted of item responses of 800 fifth-grade students who participated in the 1977 na-

tional standardization of the ITBS. The sample included 200 members of each of the following groups: black males, black females, white males, and white females. From each school building included in the study, all of which participated in the ITBS standardization, equal numbers of each of the four groups were randomly selected. Thus, the sample was not only balanced by both race and sex, but the confounding of interschool curriculum differences with ethnic group membership, common to many item bias studies, was also partially controlled.

In addition, the black students were randomly divided into two samples of 200 students each with 100 males and 100 females in each sample. The same procedure was followed for white students. Item bias indices were calculated for the first sample of black vs. the first sample of white students. Indices were also calculated for the second sample of black vs. the second sample of white students. The item bias indices were calculated separately for each of the eleven ITBS tests. Identical procedures were followed for the female vs. male comparisons.

For both race and sex comparisons, the reliability of each item bias index was investigated by correlating the values of the index across random samples of examinees. Additionally, items were classified as either biased or unbiased using the Difficulty, Delta, and Scheuneman indices. Items with Difficulty or Delta indices above .75 were classified as biased in accordance with Rudner et al. (1980b). Items with Scheuneman index values which surpassed the .05 critical value of a chi-square distribution with four degrees of freedom were classified as biased, in accordance with Scheuneman (1979). The agreement in classification of items across random samples by a given index was used as another indicator of the reliability of each of these three item bias indices.

Results

An attempt was made to estimate the item parameters for the three-parameter logistic item response model using separate LOGIST (Wood, Wingersky, & Lord, 1976) runs for each randomly equivalent sample of 400 examinees. However, LOGIST failed to converge. Because of these convergence problems, the parameter estimation was completed using all 800 examinees. The 3-Parameter indices were calculated using these parameter estimates following the same general procedures as were followed for the other indices. The use of parameter estimates from the combined sample results in a dependency between indices across randomly equivalent samples. Therefore, the reported reliabilities for the 3-Parameter index are probably overestimates of the actual values of the index. For this reason, the index was calculated for only two of the tests. The vocabulary and language usage tests were chosen because they produced the highest reliabilities for the other item bias indices.

The means and standard deviations of raw scores on each test are presented in Table 1. (The following abbreviations for test titles are used in all tables: Vocabulary (Vocab), Reading (Rdg), Spelling (Spell), Capitalization (Cap'n), Punctuation (Punct), Language Usage (Usage), Visual Materials (Vis'l), Reference Materials (Refs), Mathematics Concepts (MConc), Mathematics Problem Solving (MProb), and Mathematics Computation (MComp).) For these data, the means and standard deviations are larger for whites than for blacks; for most tests, females have slightly higher means than males.

The reliabilities of item bias indices for the race comparison are presented in Table 2. The reliabilities were generally very low to (at best) moderate, with few surpassing the .05 critical value. The reliabilities for the Language Usage test were the only ones which were consistently moderate across indices. Overall, the Difficulty and Delta indices tended to produce more reliable results than any of the other indices for the race comparison. However, Hunter (1975) has illustrated how mean differences between groups can lead to large values of these bias statistics, even when the item is not biased. Thus, the reliability of these indices may have been more an artifact of the substantial mean differences between blacks and whites than reliability for detecting item bias per se. The Scheuneman index tended to produce the least reliable results for the race comparison. For the Vocabulary and Language Usage tests, the 3-Parameter index tended to have a lower reliability than the other indices.

Table 1
Means and Standard Deviations of Raw Scores by Race and Sex

Test	No. of Items		Race		Sex		Overall
			Blacks	Whites	Females	Males	
Vocab	39	Mean	13.6	21.7	17.9	17.4	17.6
		S.D.	6.9	9.2	8.6	9.5	9.1
Rdg	54	Mean	17.5	26.3	22.2	21.5	21.9
		S.D.	7.1	11.0	9.8	10.7	10.3
Spell	40	Mean	17.5	22.2	22.0	17.9	20.0
		S.D.	8.9	9.3	9.2	9.0	9.3
Cap'n	30	Mean	12.3	15.8	15.0	13.1	14.1
		S.D.	4.7	5.7	5.3	5.6	5.5
Punct	30	Mean	10.6	14.7	13.5	11.8	12.6
		S.D.	4.6	6.2	5.8	5.6	5.8
Usage	30	Mean	9.6	15.5	13.2	12.0	12.6
		S.D.	4.7	6.8	6.4	6.6	6.5
Vis'l	46	Mean	15.6	21.8	18.6	18.8	18.7
		S.D.	5.3	7.5	6.6	7.7	7.2
Refs	45	Mean	17.6	23.8	21.9	19.5	20.7
		S.D.	7.2	9.7	8.9	9.2	9.1
MConc	37	Mean	13.0	17.7	15.6	15.0	15.3
		S.D.	5.3	6.6	6.2	6.7	6.5
MProb	27	Mean	9.5	13.1	11.3	11.4	11.3
		S.D.	4.2	5.4	4.8	5.5	5.1
MComp	45	Mean	19.8	22.3	22.2	19.9	21.1
		S.D.	7.4	8.2	7.6	8.0	7.9

The reliabilities of the item bias indices for the sex comparison are presented in Table 3. The reliabilities were generally very low. In fact, there is little evidence to suggest that the reliabilities for any index, except possibly the Scheuneman index, were above zero.

For the sake of completeness, the reliabilities of signed versions of all but the Scheuneman index were calculated for both the race and sex comparisons. The reliabilities for signed versions of the Difficulty, Delta, Biserial, and Point Biserial indices were calculated as described above except that the absolute value of the difference was not taken. The signed 3-Parameter index was the overall index described in Linn and Harnisch (1981). Although the reliabilities for the signed indices were somewhat greater than for the unsigned indices, the conclusions stated above still hold.[1]

[1]Tables for the signed indices, corresponding to Tables 2 and 3, may be obtained from the first author on request.

Additionally, the values of each item bias index were pooled over all of the items in the test battery, and the reliability of each index and the intercorrelations among indices—across randomly equivalent samples—were estimated. The intercorrelations among item bias indices across all tests for the race comparison are shown in Table 4. The diagonal entries represent the indices' reliabilities across tests. These reliabilities were fairly low. The values above the diagonal represent the average intercorrelations among indices across samples. For example, the .29 value in the table represents the average of two correlations. The first was the correlation between the Difficulty index for the first random sample and the Delta index for the second random sample. The second correlation included in the average was between the Difficulty index for the second random sample and the Delta index for the first random sample. The values above the diagonal were used in combination with the reli-

Table 2
Reliability of Item Bias Indices for Race

Test	No. of Items	Bias Index: Difficulty	Delta	Biserial	Point Biserial	Scheuneman	3-Parameter[a]
Vocab	39	.38*	.32*	.22	.43*	.06	.25[a]
Rdg	54	.25	.19	.04	.18	-.16	
Spell	40	.24	.21	.04	.08	.24	
Cap'n	30	-.09	-.07	.44*	.47*	.31	
Punct	30	.45*	.35*	.17	.24	.26	
Usage	30	.48*	.55*	.49*	.64*	.55*	.36*
Vis'l	46	.41*	.24	.07	.18	.04	
Refs	45	.01	-.07	.03	.07	.06	
MConc	37	.19	.14	.21	.14	-.30	
MProb	27	.13	.08	.29	.37*	.04	
MComp	45	-.06	-.01	.04	.09	.30	
Median		.24	.19	.17	.18	.06	--

*p < .05

[a]Index computed only for tests with values given.

Table 3
Reliability of Item Bias Indices for Sex

Test	No. of Items	Bias Index: Difficulty	Delta	Biserial	Point Biserial	Scheuneman	3-Parameter[a]
Vocab	39	.22	.22	.14	.11	.01	.09
Rdg	54	.19	.15	.08	.12	.34*	
Spell	40	-.19	-.15	-.23	-.23	.00	
Cap'n	30	-.21	-.13	-.14	-.19	.18	
Punct	30	.23	.19	-.15	-.11	.11	
Usage	30	-.11	-.14	-.05	-.03	.31	-.16
Vis'l	46	.14	.10	.21	.18	.03	
Refs	45	-.15	-.16	-.09	-.09	.08	
MConc	37	-.04	-.10	.22	.22	.37*	
MProb	27	-.17	-.12	.05	.07	-.02	
MComp	45	.10	.04	.00	-.11	.38*	
Median		-.04	-.10	.00	-.03	.11	--

*p < .05

[a]Index computed only for tests with values given.

abilities to arrive at the disattenuated correlations presented below the diagonal in Table 4.

The disattenuated correlations strongly suggest (1) that the Difficulty and Delta indices both reflect the same item bias property, (2) that the Biserial and Point Biserial indices both reflect the same item bias property, and (3) that the Difficulty and Delta indices reflect a very different item bias property than that reflected by the Biserial and Point Biserial indices. Additionally, the disattenuated correlations suggest that the Scheuneman index reflects properties reflected by both the Difficulty/Delta indices and Biserial/Point Biserial indices of item bias.

Table 5 presents the intercorrelations among bias indices for the sex comparison. The reliabilities as well as the intercorrelations among indices were negligible. Disattenuated correlations are not presented since all of the reliabilities in the table failed to surpass the .05 critical value. Overall, the results suggested little or no consistency for the sex comparison across random samples for any index.

The numbers of items classified as biased by the Difficulty, Delta, and Scheuneman indices are presented in Table 6 for the race comparison and in Table 7 for the sex comparison. Items with Difficulty or Delta indices above .75 or Scheuneman indices above the .05 critical level for a chi-square distribution with 4 degrees of freedom were classified as biased. Items with Scheuneman indices above the .20 critical value were classified as biased in a second classification for race. In these tables, the number of biased items in both samples refers to the number of items classified as biased in both Sample 1 and Sample 2. The agreement of classifications across samples was evaluated using chi-square tests of independence with Yates' correction. For the race comparison, the statistics were 4.70 for Difficulty, 3.32 for Delta, .66 for Scheuneman 1, and .63 for Scheuneman 2. Only the test for the Difficulty index surpassed the .05 critical value of the chi-square distribution with 1 degree of freedom. For the sex comparison, the statistics were 7.86 for Difficulty, 10.72 for Delta, and 2.09 for Scheuneman, with the tests for Difficulty and Delta exceeding the .05 critical value. The results presented suggest that there was, at best, minimal agreement across randomly equivalent samples.

Discussion

This investigation employed both correlation and classification consistency approaches to study the reliability of item bias statistics. The correlation statistic, as used here, represents the proportion of variability in item bias statistics that is due to the variability in item bias parameters for a test. (An item bias parameter is the expected value of an item bias statistic over repeated sampling from a

Table 4
Correlations Between Item Bias Indices Across All Tests for Race

Index	Difficulty	Delta	Biserial	Point Biserial	Scheuneman
Difficulty	.34*	.29*	.00	.01	.06
Delta	.99+	.27*	.02	.03	.07
Biserial	.01	.08	.22*	.26*	.11*
Point Biserial	.01	.11	.99	.32*	.11*
Scheuneman	.45	.36	.59	.51	.15*

* $p < .05$

Note. Underscored diagonal values are reliabilities across all tests. Values above the diagonal are average correlations between indices across all tests. Values below the diagonal are disattenuated correlations between indices across all tests. Correlations were based on 423 items.

Table 5
Correlations Between Item Bias Indices Across All Tests for Sex

Index	Difficulty	Delta	Biserial	Point Biserial	Scheuneman
Difficulty	.08	.07	.00	.01	.01
Delta		.07	.01	.03	.03
Biserial			.03	.02	.08
Point Biserial				.02	.07
Scheuneman					.07

Note. None of the correlations surpassed the .05 critical value. Underscored diagonal values are reliabilities across tests. Values above the diagonal are average correlations between indices across all tests. Correlations were based on 423 items.

Table 6
Number of Biased Items for Race
for Samples One and Two, and Both Samples

Test	No. of Items	Index											
		Difficulty			Delta			Scheuneman 1			Scheuneman 2		
		One	Two	Both	One	Two	Both	One	Two	Both	One	Two	Both
Vocab	39	3	6	0	3	5	0	1	1	0	3	3	0
Rdg	54	2	5	0	2	3	0	0	1	0	2	3	0
Spell	40	1	5	0	1	4	0	0	0	0	2	3	0
Cap'n	30	0	0	0	0	0	0	1	2	0	4	4	2
Punct	30	4	6	1	5	8	2	1	0	0	3	3	0
Usage	30	6	4	2	6	4	2	1	1	1	4	5	1
Vis'l	46	8	7	3	5	4	1	0	1	0	5	7	1
Refs	45	5	2	1	6	3	1	0	0	0	3	2	1
MConc	37	2	1	0	3	1	0	1	0	0	5	2	0
MProb	27	2	3	0	2	3	0	1	2	0	4	5	1
MComp	45	0	0	0	0	0	0	0	4	0	4	7	0
TOTAL													
Number	423	33	39	7	33	35	6	6	12	1	39	44	6
Percent		8	9	2	8	8	1	1	3	0+	9	10	1

population of examinees.) The remaining variability is due to error. As such, the correlations are a useful index of reliability of item bias statistics over all items in a test. Classification consistency is also an important consideration since decisions about item bias are often made using a particular cutoff value. However, no generally accepted guidelines exist for specifying these cutoffs. For this reason, and because there is much precedent for using correlations in the study of item bias (see Ironson & Subkoviak, 1979; Scheuneman, 1980; and Shepard, et al. 1981, for some examples), correlations were emphasized. However, both approaches suggested that the item bias indices studied were not very reliable with the present data.

Standard errors in estimating item bias statistics

Table 7
Number of Biased Items for Sex
for Samples One and Two, and Both Samples

Test	No. of Items	Index								
		Difficulty			Delta			Scheuneman		
		One	Two	Both	One	Two	Both	One	Two	Both
Vocab	39	3	3	1	3	2	1	3	0	0
Rdg	54	3	2	1	3	2	1	0	0	0
Spell	40	1	2	0	0	2	0	0	0	0
Cap'n	30	0	0	0	0	0	0	0	1	0
Punct	30	1	2	0	1	2	0	1	0	0
Usage	30	2	1	0	1	1	0	0	0	0
Vis'l	46	2	3	1	2	3	1	0	2	0
Refs	45	2	1	0	2	0	0	2	0	0
MConc	37	0	1	0	1	1	0	1	0	0
MProb	27	1	0	0	1	0	0	0	0	0
MComp	45	0	0	0	0	0	0	2	1	0
TOTAL										
Number	423	15	15	3	14	13	3	9	4	0
Percent		4	4	1	3	3	1	2	1	0

were not used in the present study because expressions for them have not yet been developed. Even if these standard errors had been available, however, they would not necessarily be comparable across different indices since each index is scaled differently. Additionally, separate standard errors would probably be required for each parameter value of a given statistic. Thus, an individual standard error probably would not be an indicator of error variability across all items in a test. As mentioned previously, however, one minus the correlation across independent random samples represents the average error variability for an index over items in a test. Since the correlation is standardized, it is comparable across indices.

One potential explanation of the instability of the indices in the present study is that there is little bias on the ITBS. Recall that item bias statistics were *not* used in the development of these forms of the ITBS. However, the items were screened for bias by experts before inclusion in final forms. If there is little bias on the ITBS, then it would appear that the expert review procedures avoid bias and, consequently, the use of item bias indices could be expected to provide little additional information about bias. If this reasoning is correct, the use of item bias statistics during the pretesting phase in this testing program is unwarranted.

The results from this study suggest that item bias indices may not lead to reliable decisions about bias. Further research is needed to ascertain the sample sizes necessary to produce sufficiently stable results to be useful. Because of the difficulty of obtaining samples of sufficient size and samples where race and interschool curriculum differences are not hopelessly confounded, such research may be virtually impossible to carry out using real subjects. One possible approach would be to do monte carlo studies in which the degree of bias is experimentally manipulated and then crossed with various sample sizes.

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–105.

Berk, R. A. (Ed.). (1982). *Handbook of methods for*

detecting test bias. Baltimore MD: The Johns Hopkins University Press.

Hunter, J. E. (1975). *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 Invitational Conference on Test Bias, Annapolis MD.

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

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* (pp. 117–160). Baltimore MD: The Johns Hopkins University Press.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

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. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement, 18,* 229–248.

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

Qualls, A., & Hoover, H. D. (1981, April). *Black and white teacher ratings of elementary achievement test items for potential race favoritism*. Paper presented at the Annual Convention of the American Educational Research Association, Los Angeles.

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

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

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

Scheuneman, J. (1980, April). *Consistency across administrations of certain indices of bias in test items*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

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

Wood, R. L., Wingersky, M. S., & Lord, R. M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76–6). Princeton NJ: Educational Testing Service.

Author's Address

Send requests for reprints and further information to H. D. Hoover, 316 Lindquist Center, The University of Iowa, Iowa City IA 52242, U.S.A.