DOCUMENT RESUME

ED 234 061 TM 830 516

AUTHOR Doolittle, Allen E.
TITLE The Reliability of Measuring Differential Item Performance.
PUB DATE Apr 83
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Black Students; *Item Analysis; *Racial Factors; *Reliability; Statistical Analysis; *Test Bias; *Test Items; White Students
IDENTIFIERS *ACT Assessment

ABSTRACT

The stability of selected indices for detecting differential item performance (item bias), from one randomly equivalent sample to another, is addressed. Some recent research has criticized these indices as too unreliable for utility in measuring bias in achievement test items. Using data from a national testing of the ACT Assessment, however, this study suggests that the reliability of the indices is situation-specific. Bias detection indices may be viewed as most reliable in testing situations that involve large sample sizes and some item heterogeneity. A preference is also stated for assessing reliability based on signed rather than unsigned indices. (Author)

The Reliability of Measuring Differential Item Performance

Allen E. Doolittle

THE AMERICAN COLLEGE TESTING PROGRAM
P.O. BOX 168
IOWA CITY, IOWA 52244

Paper presented at the annual meeting of
the American Educational Research Association

Montreal

April, 1983

ABSTRACT

The stability of selected indices for detecting differential item performance (item bias), from one randomly equivalent sample to another, is addressed. Some recent research has criticized these indices as too unreliable for utility in measuring bias in achievement test items. Using data from a national testing of the ACT Assessment, however, this study suggests that the reliability of the indices is situation-specific. Bias detection indices may be viewed as most reliable in testing situations that involve large sample sizes and some item heterogeneity. A preference is also stated for assessing reliability based on signed rather than unsigned indices.

THE RELIABILITY OF MEASURING DIFFERENTIAL ITEM PERFORMANCE

Statistical procedures for detecting differential item performance, or item bias, have been studied extensively during the past ten years (see Rudner, Getson, & Knight, 1980a; Shepard, 1981). Most of this research has addressed the very important, but elusive issue of validity for the various procedures. The reliability of item bias procedures, however, has not been thoroughly examined. After a review of the literature, Ironson (1982) suggested that more information on the reliability of various bias indices was needed.

Kolen and Hoover (1982) specifically addressed the reliability issue in their recent work. Using item responses of fifth grade students on the Iowa Tests of Basic Skills (ITBS), they found six "unsigned" procedures to have very little reliability (stability from one randomly equivalent group to another) for detecting bias with sample sizes of 200 per group. Median reliabilities across the subtests of the ITBS ranged from .06 to .24 for the various indices. Their highest reliabilities, most between .2 and .5, were found with language arts subtests.

The low reliabilities found by Kolen and Hoover, however, do not suggest that bias indices are universally unreliable. Certainly factors, such as item type, sampled population, utilized sample sizes, and type of index (signed or unsigned) may have contributed to their results. The ITBS is a well-edited battery of achievement tests, closely tied to the basic academic skills for each grade level. Conceivably, tests that are not as closely tied to specific curricula might consist of more heterogeneous items that could lead to more instances of differential item performance, greater levels of reliability and, consequently, greater potential utility for bias indices.

Sample size, too, can have a considerable effect on observed reliabilities. In their conclusion, Kolen and Hoover suggested that if bias indices are to be useful, research is needed to determine the sample sizes necessary for stable results.

Another consideration is whether signed or unsigned item bias indices (Ironson & Subkoviak, 1979) are used. Kolen and Hoover emphasized the unsigned versions in their research, "since item screening, as usually conceived, involves eliminating items biased against any group" (p. 3). This is not an illogical position from a test development standpoint. However, unsigned bias statistics do not take advantage of all available information (the direction of the "bias") as do signed indices. The result is unnecessarily low estimates of reliability.

Objectives

The major objective of this research was to examine the issue of reliable detection of item bias, as it applies to race, using:

1. a professionally-developed test, but one less closely tied to specific curricula (and more heterogeneous) than the ITBS;
2. varied sample sizes of 200 and 1000 in each group; and
3. both signed and unsigned versions of six bias indices.

A supplementary objective was to examine intercorrelations among the various indices.

Techniques

Six indices of item bias were evaluated. All of these statistics rely on internal analyses of the test to identify deviant items.

Item Difficulty Index (TID). This index is the result of the transformed item difficulties procedure (Angoff & Ford, 1973). The TID approach is based on the relative difficulty of an item for each of two groups, controlling for total test score. Items are considered biased if they are relatively more difficult for one group than for another.

Point Biserial Index (PBIS). This index is represented by the difference between the point biserial correlations, for each group, of the item with total score. It is particularly sensitive to relative group differences in item discrimination.

Item-Group Partial Correlation (IGP). The IGP index (Stricker, 1982) is the correlation of the item, scored right or wrong, and group membership, controlling for total score. This index was developed as a readily interpretable measure of differential item performance.

Modified Chi-Square Index (MOD CHI). This index is an approximate chi-square index (Scheuneman, 1979) that is based on a contingency table of correct item responses, corresponding to two groups and some finite number of score intervals (4 were used in this study). The use of matched score intervals roughly serves to equate the two groups, within an interval, on total test performance. The modified chi-square index is sensitive to group differences in both item difficulty and item discrimination.

Chi-Square Index (CHI SQR). The full chi-square index (Shepard, Camilli, & Averill, 1980) is an extension of Scheuneman's contingency table analysis of correct responses to include a similar analysis of incorrect responses. This approach is also expected to be sensitive to group differences in both item difficulty and discrimination.

3-Parameter Index (L&H). This index was proposed by Linn and Harnisch (1981) as a small sample alternative to existing 3-parameter indices that require larger sample sizes. To calculate the index, the item and ability parameters of the 3-parameter item response theory model are estimated for the total sample. The two groups are then separated. The difference is taken between each examinee's probability of correctly answering the item and the examinee's actual response to the item (1=correct; 0=incorrect). This difference is then standardized and averaged over the examinees in each group. The index is the sum of the mean values for each group (Kolen & Hoover, 1982).

Methodology

The data consisted of item responses on the 75-item English Usage subtest of the ACT Assessment (ACTE)[1] by 4000 college-bound, high school students in April, 1980. The total sample included 2000 randomly selected black (62.3% female) and 2000 randomly selected white (54.3% female) students. Mean raw score performance was 41.6 for the whites and 28.3 for the blacks. The standard deviation was 10.7 for each group. The initial and replication samples of 400 and 2000 cases each were randomly selected without replacement from this pool of students.

[1] The ACT Assessment is an achievement test directly related to high school instruction. However, since its focus is on the diverse curricula taught in high schools, it is thought to be less closely tied to curricula than tests, such as the ITBS, aimed specifically at achievement in the basic skills.

Item bias indices were calculated for each subsample. The reliability of each index was then indicated by the correlation between values of the index for the two samples of 200 black and 200 white students and for the two samples of 1000 black and 1000 white students. Since signed and unsigned versions of the indices were investigated, each index has four reliabilities associated with it: the signed and unsigned versions for the 400-case samples and the signed and unsigned versions for the 2000-case samples.

An additional approach to index reliability was performed, based only on the specific items that were identified as most biased. This approach was useful because it provided a measure of the practical reliability of each procedure for identifying deviant items. The ten most deviant items for each sample, as determined by each procedure, were identified (the ten with the greatest absolute magnitude of the index). Unweighted Kappa coefficients (Cohen, 1960) were then calculated for each procedure and each pair of samples. Since items were identified on the basis of the absolute magnitude of the indices, this approach to reliability is closely related to the unsigned correlational reliabilities.

As with reliability, the interrelationships of the indices were examined in two ways. First, by the intercorrelations of the signed indices, obtained for one of the 2000-case samples; and secondly, by Kappa coefficients of agreement between items selected as most deviant by each index.

Results

The reliabilities of the small and the large sample bias indices are shown in Table 1. As expected, in every case the reliabilities were higher for the larger than for the smaller samples. Also, as expected, the reliabilities for all the signed indices were higher than for their unsigned counterparts. Regardless of whether the signed or unsigned versions of the bias indices are compared, though, the TID approach seemed to be the most reliable. However, as Hunter (1975) has pointed out, this index is spuriously sensitive to group differences in performance. The relatively high degree of reliability for this procedure may be an artifact of the substantial performance difference between blacks and whites on the items (Kolen & Hoover, 1982). The remaining indices seemed to be about equal in reliability.

Table 2 presents the Kappa coefficients and the number of deviant items selected in common, between samples, by each index. It should be noted that separate consideration of signed and unsigned versions of the indices is not presented here, because the results would be the same. That is, the same sets of items, selected on the basis of the absolute magnitude of the index, would result. However, this

TABLE 1

Reliabilities of Signed and Unsigned Bias Indices*

	Small Samples (N = 400)		Large Samples (N = 2000)	
Index	Signed	Unsigned	Signed	Unsigned
TID	.72	.39	.91	.80
PBIS	.59	.48	.66	.58
IGP	.60	.31	.74	.50
MOD CHI	.62	.47	.77	.69
CHI SQR	.58	.32	.76	.66
L&H	.61	.22	.79	.51

* All values in the table are statistically significant (p < .001).

analysis is most closely related to the reliability estimation of the unsigned versions of the indices. At least for the large samples, the TID procedure again seemed to produce the most reliable results. Eight of the ten deviant items (80%), identified by the TID procedure using the first of the large samples, were also identified using the second large sample. About 50 percent agreement between samples was evident for the other bias measures.

From a pure measurement perspective, the signed bias indices are preferred to the unsigned versions since they reflect not only magnitude but directionality as well. A better understanding of the relationships between the bias indices is thus found by investigating the signed versions. Table 3 shows the intercorrelations among the signed bias indices for one of the 2000-case samples. Table 4 presents

TABLE 2

Consistency of Deviant Item Selection

	Small Samples (N = 400)		Large Samples (N = 2000)	
Index	Kappa	Common items*	Kappa	Common items
TID	.31	4	.77	8
PBIS	.54	6	.54	6
IGP	.14	3	.31	4
MOD CHI	.31	4	.42	5
CHI SQR	.31	4	.42	5
L&H	.14	3	.31	4

* The number of items that are common to each sample's set of ten most deviant items.

the same intercorrelation matrix after correcting for attenuation. The results indicate a great deal of similarity among the IGP, the modified Chi-square, the full Chi-square, and the Linn & Harnisch measures. The TID procedure seems to be moderately related to these procedures, while the Point Biserial approach seems to stand alone. These results are consistent with expectations. The Point Biserial index is the only measure to emphasize group differences in item discrimination and it clearly does not correlate positively with the other procedures. The TID index emphasizes only group differences in item difficulty, and it, too, seems to stand at least somewhat apart from the others. The remaining four indices are sensitive to group differences in item difficulty and discrimination, and they seem to produce very similar results.

TABLE 3

Intercorrelations of Signed Bias Indices*

	TID	PBIS	IGP	MOD CHI	CHI SQR	L&H
TID	1.00	-.03	.51	.45	.61	.53
PBIS		1.00	-.13	-.11	.07	.03
IGP			1.00	.67	.92	.95
MOD CHI				1.00	.70	.72
CHI SQR					1.00	.94
L&H						1.00

* Correlations based on one of the 2000-case samples and the 75 ACTE items.

TABLE 4

Intercorrelations of Unattenuated Signed Bias Indices

	TID	PBIS	IGP	MOD CHI	CHI SQR	L&H
TID	1.00	-.04	.62	.54	.73	.63
PBIS		1.00	-.19	-.15	.10	.04
IGP			1.00	.89	.99	.99
MOD CHI				1.00	.92	.92
CHI SQR					1.00	.99
L&H						1.00

Table 5 presents measures of agreement in selection of deviant items between the different bias indices. As shown previously in Table 4, the TID and Point Biserial indices tend to stand apart, whereas the other four indices seem to be more closely related. To illustrate the commonality between the IGP, the modified Chi-square, the full Chi-square, and the Linn & Harnisch indices, four items were identified as being among the ten most deviant items by all four procedures. Another four items were identified by three of these four procedures, using one of the 2000-case samples.

TABLE 5

Measures of Agreement in Deviant Item Selection*

	TID	PBIS	IGP	MOD CHI	CHI SQR	L&H
TID	1.00	-.15(0)	.31(4)	-.04(1)	.19(3)	.08(2)
PBIS		1.00	-.04(1)	.08(2)	-.04(1)	.08(2)
IGP			1.00	.54(6)	.42(5)	.77(8)
MOD CHI				1.00	.54(6)	.77(8)
CHI SQR					1.00	.54(6)
L&H						1.00

* The first number for each combination is the Kappa coefficient of agreement. The values in parentheses are the numbers of items common to the set of 10 most deviant items selected by each procedure, using one of the 2000-case samples.

Discussion

The results demonstrate that testing situations do exist in which bias indices can reliably detect differential item performance. Like most phenomena in the social sciences, though, the reliability of bias indices seems to be situation-specific. Kolen and Hoover (1982) effectively argued that statistical bias detection procedures were not very reliable and, consequently, not very useful within the current test development process of the ITBS. However, with more heterogeneous tests, such as the ACT Assessment, and with larger sample sizes, commonly investigated bias indices can attain potentially useful levels of reliability.

The manner in which the bias indices are to be used is also important. If they are used in the process of learning more about the functioning of various items or item types for different groups, both the degree and directionality of bias are important. The relevant reliability analysis for this use of item bias indices is correlational, as shown in Table 1. Particularly with the larger samples, but also with the 400-case samples, the signed indices seem to be reasonably reliable for this purpose.

If bias indices are used as screening devices in the test development process, the relevant analysis is the stability of item classification. Although Table 2 indicates some commonality, these data do not suggest that the bias indices can be relied on to the exclusion of expert editorial review. In fact, with some curriculum-bound tests and a test development process that includes several stages of thorough editorial review, these indices may be relatively useless (Kolen & Hoover, 1982). The indices seem to offer more promise, however, when used with more heterogeneous tests and when used as a tool to screen items for more extensive editorial review. Neither common sense nor the results of this study suggest that bias indices should supercede expert judgment on the desirability of an item within a test.

Future efforts in studying the reliability of item bias procedures might focus on other indices, or the combination of two or more indices. The investigation of procedures stemming from latent trait theory (Ironson, 1982; Lord, 1980), for instance, would certainly be in order. The joint reliability of two or more, relatively independent and easily computed indices (such as the TID and PBIS) might also be useful.

Finally, Monte Carlo studies could be very useful in the systematic exploration of index reliability. Use of simulated data, a la Rudner, Getson, and Knight (1980b) in validity research, might help clarify the effects of different types of tests and examinee populations on the reliability of statistical item bias procedures.

REFERENCES

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

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, 20(1), 37-46.

Hunter, J. E. *A critical analysis of the use of item means and ite-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, 1975.

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

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

Kolen, M.J., & Hoover, H.D. The reliability of selected item bias procedures. Paper presented at the Annual Meeting of the American Educational Research Association, New York, March, 1982.

Linn, R.L., & Harnisch, D.L. Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 1981,18(3), 109-118.

Lord, F. M. *Applications of item response theory to practical testing problems*. *Hillsdale, NJ: Erlbaum, 1980*.

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

Rudner, L. M., Getson, P. R., & Knight, D. L. A Monte Carlo comparison of several methods of assessing item bias. *Journal of Educational Measurement*, 1980, 17, 1-10.

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

Shepard, L. A. Identifying bias in test items. In B. F. Green (Ed.), New Directions for Testing and Measurement: Issues in Testing -- Coaching, Disclosure, and Ethnic Bias, no. 11, San Francisco: Jossey-Bass, 1981.

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

Stricker, L.J. Identifying test items that perform differentially in population subgroups: a partial correlation index. Applied Psychological Measurement, 1982, 6(3), 261-273.