Effect of Sample Size, Number of Biased Items, and Magnitude of Bias on a Two-Stage Item Bias Estimation Method

M. David Miller, University of Florida

T. C. Oshima, Georgia State University

A two-stage procedure for estimating item bias was examined with six indexes of item bias and with the Mantel-Haenszel (MH) statistic; the sample size, the number of biased items, and the magnitude of the bias were varied. The second stage of the procedure did not identify substantial numbers of false positives (unbiased items identified as biased). However, the identification of true positives in the second stage was useful only when the magnitude of the bias was not small and the number of biased items was large (20% or 40% of the test). The weighted indexes tended to identify more true and false positives than their unweighted item response theory counterparts. Finally, the MH statistic identified fewer false positives, but did not identify small bias as well as the item response theory indexes. *Index terms: differential item functioning, item bias, Mantel-Haenszel statistic, two-stage bias estimation.*

Studies of item bias or differential item functioning have shown that statistical artifacts can lead to the identification of unbiased items as biased and the underidentification of biased items (e.g., Shepard, Camilli, & Williams, 1984). Biased items may be misidentified using item response theory (IRT) because the biased items are included in the initial estimation of item parameters. These biased items could potentially introduce some second trait into the unidimensional estimates of the trait and item parameters. Thus, the initial unidimensional parameter estimates, which form the basis for the estimation of bias, may be a weighted composite of two or more traits (Ackerman, 1988; Wang, 1986). At a minimum, the unidimensional estimate is affected by some differential variability on the nuisance dimension. Because of this concern, Marco (1977) and Lord (1980), among others, suggested using a multistage estimation procedure to identify biased items using IRT. Lord pointed out that when large numbers of items are biased, the test may not be strictly unidimensional. As a result, interpretation of the trait estimates (θs) obtained for the two groups may not be comparable. Thus, a two-stage estimation procedure may be necessary to derive a unidimensional trait to estimate bias.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 4, December 1992, pp. 381-388
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/040381-08$1.65

Lord's two-stage, or "purification," procedure is conducted as follows:

1. Initially estimate bias;
2. Remove the biased items and reestimate θ for both groups combined; and
3. With fixed θs, reestimate bias for the total test.

This two-stage procedure makes the estimation of θ and the equating of groups onto a common scale independent of the effects of potentially biased items identified in Stage 1. However, the reestimation aspect of the two-stage procedure (Steps 2 and 3) has not been applied in practice nor in studies of item bias.

Some researchers have examined fully iterative procedures in the analysis of item bias. For example, several non-IRT indexes of item bias are based on multistage estimation procedures, such as the iterative logit method (Kok, Mellenbergh, & Van der Flier, 1985; Van der Flier, Mellenbergh, Ader, & Wijn, 1984). In addition, a two-stage version of the Mantel-Haenszel (MH) statistic has been proposed by Holland & Thayer (1988).

Several researchers have examined an iterative procedure for estimating item bias that focuses on the equating or linking of groups using IRT (Candell & Drasgow, 1988; Drasgow, 1987; Lautenschlager & Park, 1988; Park & Lautenschlager, 1990). For example, Candell & Drasgow examined a variation of the Lord (1980) two-stage procedure. They emphasized the need for a common metric to assess item bias and the need to examine linking, or equating, errors (Shepard et al., 1984) that can lead to the spurious identification of biased items. Candell & Drasgow's iterative procedure, which does not require the reestimation of IRT parameters, is implemented as follows:

1. Initially estimate item bias;
2. Relink, or equate, items across groups without the items identified as biased;
3. Reestimate item bias; and
4. Continue Steps 2 and 3 iteratively until no difference in bias is found on two consecutive passes.

Candell & Drasgow found that iteratively linking items improved the identification of biased items in a simulation study.

However, relinking may be insufficient if the biased items create a multidimensional test, as suggested by Lord (1980). The multidimensionality will have different effects on the item parameters in the two groups when their distributions and correlations of the traits differ (Ackerman, 1988, 1992; Oshima & Miller, 1990; Wang, 1986). Thus, Park & Lautenschlager (1990) examined an iterative item bias estimation procedure that included both relinking and trait scale purification. They found this procedure to be more effective than iterative linking alone.

A variation of the trait scale purification procedure was examined here with systematic variations in sample size, magnitude of bias, and number of biased items. Because the final results of Park & Lautenschlager (1990) were often similar to the results after a single iteration (no further iterations in two of four conditions and only one additional item was identified in the two conditions that had more than one iteration), a two-stage bias estimation procedure was used here. A parallel two-stage estimation procedure was conducted using the MH statistic, which requires less computing time and cost.

Method

Simulation

Items. Item parameters for a 40-item test were generated for the two-parameter logistic model (2PLM). The 2PLM was selected because the recovery of the lower asymptote in the three-parameter logistic model is poor (Hulin, Lissak, & Drasgow, 1982), and poor estimation of an item parameter would introduce a confounding factor that could affect bias estimation.

A unidimensional model was used to simulate data, and bias was introduced by altering the item difficulty (b) parameter in the bias group as has previously been done (e.g., Shepard, Camilli, & Williams, 1985). An alternative approach to introducing bias is to use a multidimensional IRT model. However, no agreed on criteria for simulating bias have been established in the multidimensional case. Ackerman (1992) pointed out that several different parameter shifts can lead to bias (e.g., difficulty or θ parameters). In addition, Oshima & Miller (1992) showed that traditional bias indexes may not be sensitive enough to some multidimensional parameter shifts, depending on the relative importance of the secondary trait.

Using SAS (1985), the item discrimination (a) parameters were generated to be log-normal and the b parameters were generated to be normal [N(0, 1)]. The resulting a distribution had a mean of 1.13 with a standard deviation (SD) of .63 (a ranged from .41 to 3.82). The b distribution had a mean of .25 with a SD of 1.05 (range from -1.99 to 2.10).

Simulated data. The θ distributions were randomly generated from a [N(0, 1)] distribution.

The individual item responses were generated using the probability of a correct response from the 2PLM and a random uniform distribution. A correct response was recorded when a random number taken from a uniform distribution between 0 and 1 was less than or equal to the probability of a correct response; otherwise, the response to the item was considered to be incorrect.

Conditions. A total of 26 datasets (two baseline and 24 bias conditions) of item responses were generated. To establish a baseline estimate of differences in the item response functions (IRFs), two baseline samples of 1,000 simulees were generated using the item parameters and simulation procedures described above. In addition, 24 bias conditions were simulated to examine the two-stage item bias estimation procedure. The 24 bias conditions were based on a completely crossed design (2 × 3 × 4)—the number of simulees (1,000 in each group or 1,000 in one group and 300 in the other group), the magnitude of bias (small, moderate, or mixed), and the number of biased items (5%, 10%, 20%, and 40%) were varied in the bias samples.

The number of simulees was varied to examine the effect of group size on item bias identification; that is, to examine effects when both the baseline and bias group were equally large (e.g., $N_1 = N_2 = 1{,}000$) versus when one group (e.g., minorities) was substantially smaller (e.g, the $N_1 = 1{,}000$ and the $N_2 = 300$ cases). The magnitude of bias was specified as in Shepard et al. (1985), who found that the average bias found in an arithmetic test was a difference of approximately .35 in the b parameter, and the minimum detectable difference was a difference of .20. Thus, in the small bias condition, b was increased by .20. In the moderate bias condition, b in the bias group was increased by .35. In the mixed bias condition, half of the biased items were assigned a small bias and half were assigned a moderate bias.

Bias was introduced into the b parameters, and in only one direction, for two reasons: (1) to be comparable to other simulation studies of item bias (e.g., Candell & Drasgow, 1988; Shepard et al., 1985); and (2) because the cumulative effect of item bias on the test response functions for the subpopulations can be nonexistent when different items favor different groups or when the items differ only on the discrimination parameters (Stout & Shealy, 1991). As a result, the consistency in the direction of the bias for this study modeled the pervasive bias situation in which the test response functions also are expected to differ by subpopulation. The number of biased items—5%, 10%, 20%, or 40% of the 40-item test—ranged from what is typically found in studies of ethnic, race, or gender bias (5% to 10%) to the number of biased items found in studies of instructional effects (20% to 40%; Miller & Linn, 1988).

Analysis

PC-BILOG (Mislevy & Bock, 1985) was used to estimate the 2PLM for each of the 26 datasets using all the simulees and the default priors (i.e., a was log-normal; b and θ were normal). A baseline estimate of sampling fluctuations in the item bias indexes was established using the two baseline samples.

After linking the two sets of item parameters by the mean and sigma method (Hambleton & Swaminathan, 1985, p. 207), six indexes of differential item functioning based on IRT and the MH statistic were calculated for each item. The MH statistic was conditioned on the biased item and the remaining unbiased items (Ackerman, 1992). The six IRT indexes were: the signed area (SA) and unsigned area (UA) between the IRFs (Rudner, Getson, & Knight, 1980); the signed and unsigned sums of squares (SSOS and USOS, respectively; Linn, Levine, Hastings, & Wardrop, 1981); and the weighted signed and weighted unsigned sums of squares (WSSOS and WUSOS, respectively; Linn et al., 1981). The baseline estimates of the bias statistics were used to decide when an item was biased. An item was identified as biased if it exceeded the mean plus two SDs from the distribution of indexes in the baseline comparison.

In Stage 1 of the bias estimation procedure, each of the 24 bias conditions was compared to the first baseline group. That is, the bias condition item parameters were equated to the first baseline condition and bias statistics for each item were calculated. In Stage 2, the items identified as biased in Stage 1 were removed, and the bias estimation procedure was repeated with the reduced item datasets. That is, for each bias condition, the biased items were removed from the bias condition and from the first baseline condition to model what would be done in practice when items are multidimensional. The statistics from the baseline condition based on a reduced set of items would only affect the results if the number of items affects the estimation of the baseline distribution, because no multidimensionality or bias exists in the baseline. Next, the item parameters for the reduced tests were reestimated with PC-BILOG, the two sets of item parameters were equated, and the bias statistics were calculated to identify biased items in Stage 2 with the same criteria.

Results

Table 1 reports the results of the comparison of the IRFs for the two baseline conditions on the six IRT indexes. Using the bias identification criterion in Table 1, the numbers of biased items identified as biased (true positives) are reported in Table 2 for both stages of the estimation. In addition, items that were identified as biased, but were not (false positives), are reported in Table 3.

Although no consistent effect due to sample size can be seen across conditions in the identification of biased items, identification was better for the larger number of simulees when the number of biased items was small (5% or 10%) and the magnitude of the bias was small. Interestingly, these conditions approximate typical studies of bias when care has been exercised in the initial item/test construction phases. On the other hand, a large and consistent effect across conditions can be seen in the overidentification of unbiased items (Table 3). When the bias group contained 300 simulees, substantial numbers of false positives were evident across magnitude of bias, number of biased items, and the six IRT indexes of bias—averaging 5.8 false positives in Stage 1. In contrast, the average number of false positives when $N_1 = N_2 = 1{,}000$ was 1.8 across conditions and indexes in Stage 1. Although fewer false positives were identified in Stage 2, the overall trend held. That is, fewer items were identified when the size of the bias group was larger (averaging .5 item) than when the bias group was smaller (averaging 1.1). This suggests that the IRFs were less stable in the bias conditions with the smaller sample size, which resulted in the identification of larger numbers of false positive items. One notable exception to this trend was the MH statistic, which had far fewer false positives.

The effect of the magnitude of the bias was also substantial. In selecting a difference in the *b* parameters of .2, Shepard et al. (1985) modeled "the smallest detectable difference for items found biased" (p. 87) on a mathematics achievement test. Consistent with their results, fewer biased items were identified when the bias was small (Table 2). With a moderate bias (.35), a substantially higher proportion of the biased items was identified, and the mixed bias conditions fell between the small and moderate bias conditions, as expected. In addition, in both the moderate and the mixed bias conditions, 100% of the moderately biased items were detected when the number of biased items was not large (5% or 10%) and the number of simulees was large ($N_1 = N_2 = 1{,}000$). The trend from the small to moderate bias conditions suggests that

Table 1
Mean and SD of Bias Indexes for Baseline Group and Value of the Bias Identification Criterion

Index	Mean	SD	Criterion (Mean + 2 SD)
SA	0.00	.10	±.21
UA	.11	.07	.25
SSOS	.07	.88	±1.83
USOS	.56	.77	2.09
WSSOS	.04	.94	±1.91
WUSOS	.68	.77	2.23

Table 2
Number of Biased Items Exceeding the Baseline Criterion (True Positives) After Stage 1 and Number Identified in Stage 2 (In Parentheses) for Each Bias Index by Degree and Percent of Bias, for Datasets of Equal and Unequal Size

Bias and Percent of Biased Items	Item Bias Index						
	SA	UA	SSOS	USOS	WSSOS	WUSOS	MH
Equal Sample Size ($N_1 = N_2 = 1{,}000$)							
Small							
5%	0	0	1	0	1	1	1
10%	4 (1)	2	2	2	4	4	4 (1)
20%	2	1	2	1	5 (1)	5 (2)	2 (1)
40%	6 (3)	2	4 (1)	3	14 (4)	11 (4)	6 (1)
Moderate							
5%	2	2	2	2	2	2	2
10%	4	4	4	4	4	4	4
20%	7 (1)	4	7 (1)	6 (2)	6	6	7 (1)
40%	13 (2)	10 (5)	13 (3)	11 (4)	14 (1)	13	13
Mixed							
5%	1	1	1	1	2	2	2
10%	3	2	3 (1)	2	4	4	4 (1)
20%	6 (1)	3 (1)	6 (3)	2	7	7	7 (2)
40%	11 (4)	10 (2)	9 (2)	9 (2)	15 (5)	15 (5)	9 (2)
Unequal Sample Size ($N_1 = 1{,}000$; $N_2 = 300$)							
Small							
5%	0	0	0	0	0	0	0
10%	1	1	1	1	1	1	0
20%	5 (1)	5 (2)	5 (1)	5 (1)	5	6	2
40%	8 (1)	4 (1)	6	6 (2)	12 (2)	10	2
Moderate							
5%	2	2	2	2	2	2	2
10%	4	3	4	4 (1)	3	4 (1)	4
20%	6	6	7	5	7	7	2
40%	15 (5)	12 (1)	13 (3)	14 (3)	16 (1)	16 (3)	13 (5)
Mixed							
5%	2	2	2	2	2	2	2 (1)
10%	4	3	3	3	4	4	3 (1)
20%	5	4	7 (2)	6 (1)	5	5	3 (1)
40%	10 (2)	6	9	8 (1)	10 (4)	9 (3)	8 (1)

the proportion of true positives would probably increase with an even larger bias introduced, such as the *b* difference of .75 used in Candell & Drasgow (1988).

The effect of different proportions of biased items interacted with the magnitude of the bias in the identification of biased items. For the larger sample size, moderate bias, and 5% or 10% of biased items, the hit rate for detecting the biased items was 100%. The hit rate for larger numbers of biased items (20% or 40%) was not as high, but it was still substantial with moderate bias. With the small bias, the identification of bias was uniformly low across the number of true biased items.

The two-stage bias estimation procedure had a slight positive effect. When the number of biased items was large (20% or 40%), several additional items were correctly identified as biased in Stage 2. However, when there were fewer biased items with a moderate magnitude, Stage 2 did not have the potential to identify additional items because the true positives were usually identified in Stage 1. On the other hand, there

Table 3
Number of Unbiased Items Exceeding the Baseline Criterion (False Positives) After Stage 1 and Number Identified in Stage 2 (In Parentheses) for Each Bias Index by Degree and Percent of Bias, for Datasets of Equal and Unequal Size

Bias and Percent of Biased Items	Item Bias Index						
	SA	UA	SSOS	USOS	WSSOS	WUSOS	MH
Equal Sample Size ($N_1 = N_2 = 1{,}000$)							
Small							
5%	3 (1)	2 (1)	0	1	3	3	2 (1)
10%	1	2 (2)	0	0	3 (2)	1	1
20%	0	0	1 (1)	0	2 (1)	1 (1)	1
40%	2	2 (1)	2 (1)	1	5 (3)	4	3
Moderate							
5%	2 (1)	1	0	0	2 (2)	1	1
10%	1	1	1	1	2 (1)	2	0
20%	7	5 (1)	4	1	9	8	2
40%	5	2	3	1	10	10 (1)	3
Mixed							
5%	3	1	1	0	2 (1)	2	1
10%	2 (1)	0	1 (6)	0	3 (2)	3 (2)	0
20%	1	1	2 (1)	1 (1)	4 (4)	3 (2)	0
40%	3 (1)	1	3 (1)	1	6 (1)	6 (1)	2
Unequal Sample Size ($N_1 = 1{,}000$; $N_2 = 300$)							
Small							
5%	7 (1)	6	6	5	8	6	1
10%	5 (1)	5	7 (1)	6 (1)	5 (2)	3 (1)	0
20%	7	6 (2)	6	4	13 (6)	9 (1)	1
40%	2 (1)	2	1	1	6	4	0
Moderate							
5%	6 (1)	7 (1)	7 (1)	5	9	9 (1)	7 (3)
10%	2 (1)	3 (2)	3	3	10 (8)	3 (1)	2 (1)
20%	4	5 (1)	5 (1)	4	5 (1)	6 (1)	1
40%	7 (1)	8	7	6	11	10 (1)	1
Mixed							
5%	6 (2)	5	5	5	10 (2)	9 (1)	2 (2)
10%	4 (1)	6 (2)	4	2	10 (7)	9 (5)	0
20%	8	11 (2)	5	5	17 (6)	14 (3)	0
40%	4	5 (1)	5 (1)	2	13 (3)	9 (2)	0

was not a negative consequence of using a second stage to the estimation, because few false positives were identified during Stage 2 across the 24 conditions. Thus, many of the items that were identified during Stage 2 were truly biased.

The index used to identify bias also had an effect on the results. The weighted indexes (WSSOS and WUSOS) identified more biased items initially, but after Stage 2 the differences across indexes were small. During Stage 1, identification of larger numbers of biased items was related to the identification of more false positives, with the weighted indexes (WSSOS and WUSOS) identifying more false positives. The MH statistic identified moderate bias as well as the IRT indexes, and identified fewer false positives across all conditions. On the other hand, identification of items with a small bias was not as good with the MH statistic.

Conclusions

The two-stage bias estimation procedure did not have a substantial impact on the identification of biased items when the number of biased items was small (5% or 10%). It had some impact when the number of biased items was 20% of the test or more and the magnitude of the bias was not small. By contrast, the Candell & Drasgow (1988) results indicated that their iterative bias estimation procedure resulted in better identification of bias than the comparable noniterative estimation procedures under all conditions. Consistent with the results presented here, they had items with a large bias (differences in the *b*s of .75) and a substantial number of biased items (10% or 30%). Park & Lautenschlager (1990) also used conditions in which large numbers of items were biased. Although large numbers of biased items have been found in the instructional effects literature (e.g., Miller & Linn, 1988), fewer items are found in studies of racial, ethnic, or gender bias (e.g., Linn, et al., 1981). Therefore, the two-stage identification of bias may be less useful than single-stage procedures in the identification of ethnic, race, or gender bias, when the number of biased items is small and the magnitude of the bias is weak.

References

Ackerman, T. A. (1988, April). *An explanation of differential item functioning from a multidimensional perspective.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6,* 249–260.

Kok, F. G., Mellenbergh, G. J., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22,* 295–303.

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12,* 365–376.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of 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.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139–160.

Miller, M. D., & Linn, R. L. (1988). Invariance of item parameters with variations in instructional coverage. *Journal of Educational Measurement, 25,* 205–219.

Mislevy, R. J., & Bock, R. D. (1985). *PC-BILOG Version 1.1: Maximum likelihood item analysis and test scoring: Logistic model* [Computer program and manual]. Mooresville IN: Scientific Software.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indices: The effect of between group variation in trait correlations. *Journal of Educational Measurement, 27,* 273–283.

Oshima, T. C., & Miller, M. D. (1992). Item bias and multidimensionality. *Applied Psychological Measurement, 16,* 237–248.

Park, D., & Lautenschlager, G. J. (1990). Iterative linking and ability scale purification as means for improving IRT item bias detection. *Applied Psychological Measurement, 14,* 163–173.

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

SAS. (1985). *SAS User's Guide: Basic: Version 5* [Computer program and manual]. Cary NC: Author.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77–105.

Stout, W. F., & Shealy, R. (1991). *An IRT-based statistical procedure to detect test bias simultaneously present in several items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21,* 131–145.

Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at the Office of Naval Research Contractor's Conference, Galinburg TN.

Author's Address

Send requests for reprints or further information to M. David Miller, 1403 Norman Hall, University of Florida, Gainesville FL 32611, U.S.A.