DOCUMENT RESUME

ED 339 746 TM 017 678

AUTHOR Kromrey, Jeffrey D.; Parshall, Cynthia G.

TITLE Screening Items for Bias: An Empirical Comparison of the Performance of Three Indices in Small Samples of Examinees.

SPONS AGENCY Florida State Dept. of Education, Tallahassee.; University of South Florida, Tampa. Inst. for Instructional Research and Practice.

PUB DATE Nov 91

NOTE 26p.; Paper presented at the Annual Meeting of the Florida Educational Research Association (Clearwater, FL, November 13-16, 1991).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Comparative Analysis; Computer Simulation; Elementary School Teachers; Elementary Secondary Education; Higher Education; *Item Bias; Licensing Examinations (Professions); Mathematical Models; Monte Carlo Methods; *Sample Size; Secondary School Teachers; Teacher Certification; *Test Items

IDENTIFIERS *Delta Method; Empirical Research; *Mantel Haenszel Procedure; Standardization; Subject Content Knowledge

ABSTRACT

A Monte Carlo study was conducted to compare the performance of three statistical indices of test item bias in small samples of examinees. The statistical indices compared were the Delta method, the Mantel-Haenszel (MH) method, and the Standardization method. Sample sizes of 50, 100, and 200 were examined. One thousand samples of each size were drawn with replacement from each of three archival data files from three teacher subject area tests (in the areas of elementary education, early childhood education, and specific learning disabilities). Each sample was drawn so that 80% of the examinees were sampled from a reference group and 20% were sampled from a focal group. Item bias was experimentally controlled in the study, and the effectiveness of the indices was evaluated as the proportion of such biased items appropriately identified. Previous research suggesting that item bias indices such as the MH and Standardization methods should only be applied to large samples may have been overly conservative. Results support the use of statistical screening for item bias, even with samples as small as 50 examinees, and with only 10 focal group members in each sample. The MH is the best performer of these three indices, although both the MH and Standardization methods are preferable to the Delta method. Three tables present simulation data. An 11-item list of references and 3 tables are included. (SLD)

Screening Items for Bias: An Empirical Comparison of the Performance of Three Indices in Small Samples of Examinees

Jeffrey D. Kromrey

Cynthia G. Parshall

Department of Educational Measurement and Research
University of South Florida

Paper Presented at the Annual Meeting of the Florida Educational Research Association, November 13-16, 1991, Clearwater Beach, Florida

Abstract

A Monte Carlo study was conducted to compare the performance of three statistical indices of test item bias in small samples of examinees. The statistical indices compared were the Delta method, the Mantel-Haenszel method, and the Standardization method. Sample sizes of 50, 100, and 200 were examined. One thousand samples of each size were drawn with replacement from each of three archival data files from teacher subject area tests. Each sample was drawn so that 80% of the examinees were sampled from a reference group and 20% from a focal group. Item bias was experimentally controlled in the study, and the effectiveness of the indices was evaluated as the proportion of such biased items appropriately identified.

Screening Items for Bias: An Empirical Comparison of the Performance of Three Indices in Small Samples of Examinees

The evaluation of test items for bias is an important component in the analysis of examination results. Several statistical indices have been proposed as useful screening tools for test item bias (Angoff & Ford, 1973; Dorans & Kulik, 1986; Holland & Thayer, 1988). Practical research related to the utility of such indices and comparisons among the indices have been presented by Hills (1989), Perlman, Bezrucsko, Junker, Reynolds, Rice, & Schulz (1988), and Shepard, Camilli, and Williams (1985).

A practical issue which has not been adequately researched is the utility of statistical screening indices when the number of examinees is small. The purposes of this research were (a) to appraise the sensitivity of three indices of item bias in small sample situations, (b) to estimate the stability of the indices, and (c) to provide recommendations on the appropriate use of statistical indices for item bias screening in small sample testing programs.

Statistical Indices of Item Bias

Three indices of item bias were compared in this study: Angoff's Delta method (Angoff & Ford, 1973), the Mantel-Haenszel method (Holland & Thayer, 1988), and the Standardization method (Dorans & Kulik, 1986). Each index is briefly described below. More detailed treatments of these statistical indices are provided in Scheuneman and Bleistein (1989).

Angoff's Delta Method

The Delta method is based on differences in transformed item difficulty indices between the reference group (e.g., white examinees) and the focal group of examinees (e.g., black examinees). Item difficulties are computed separately for the two groups. To convert the difficulty indices to an equal interval scale, the difficulty indices are transformed to inverse standard normal deviates (z-scores). By convention, the Delta values are obtained by linearly transforming the normal deviates to a scale yielding a mean of 13 and a standard deviation of 4 (Delta = 4z + 13).

A scatterplot of the Delta values for the two groups produces an ellipse and an item is flagged as potentially biased when the coordinates of the item in such a scatterplot are distant from the majority of item points in the bivariate space. Specifically, the perpendicular distance of each item from the major axis of the ellipse is calculated, and this distance measure is used as an index of potential bias.

Mantel-Haenszel Method

The Mantel-Haenszel method provides a comparison of the performance of examinees in the reference and focal groups, while matching examinees on total test score. The method is best conceptualized by considering a series of 2 X 2 contingency tables, each table representing examinees who have received the same total score on the test:

Table for Examinees with Total Score = s			
	Item Performance		
Examinee Group	Right	Wrong	Total
Focal (f)	R_{fs}	W_{fs}	N_{fs}
Reference (r)	R_{rs}	W_{rs}	N_{rs}
Total	R_{ts}	W_{ts}	N_{ts}

The index of potential item bias is given by the weighted sum of odds ratios across the set of contingency tables:

$$\alpha_{MH} = \frac{\sum_{s} \{(R_{rs} \; W_{fs}) \; / \; N_{ts}\}}{\sum_{s} \{(R_{fs} \; W_{rs}) \; / \; N_{ts}\}}$$

With small samples of examinees, matching by total score is not feasible (such matching yields contingency tables in which most cells are empty). The procedure was modified for the small sample situation in this research by dividing the examinees into only two strata (high and low total scores). The point of demarcation between strata was computed as the median total score of the sample of focal group members.

The Standardization Method

The Standardization method provides an index based on the weighted difference in difficulty indices for the reference

and focal groups, while matching examinees on the basis of total test score:

$$D_{std} = \sum_{s} N_{fs} p_{fs} - \sum_{s} N_{fs} p_{rs}$$

where p_{fs} and p_{rs} are the item difficulty indices for the focal and reference group examinees at total score level s.

As with the modification described above for the Mantel-Haenszel procedure, the examinees in each sample were divided into only two matched groups (high and low total scores), based on the median total score of the focal group members in each sample.

Applications of Item Bias Indices to Small Samples

Dorans (1989), and Shepard et al. (1985) recommended additional research on applications of item bias indices in small samples of examinees. In an attempt to address the small sample issue, Shepard et al. (1985) examined statistical indices with focal group sizes of 300 examinees. However, even this number is considerably larger than the size of the focal group examinee samples in many small testing programs.

Hills (1989) stated that Angoff's Delta method is commonly used with small samples. The Delta method has been recommended as the best, and sometimes as the only, choice when screening examination data from small samples (Scheuneman & Bleistein, 1989). In an early study of the Delta method, Angoff and Ford (1973) evaluated 10 samples presenting numbers of cases ranging from 125 to 340.

However, problems with the Delta method and evidence of spurious bias identification led Shepard et al. (1985) to recommend that the unmodified Delta method not be used at all, even under small sample constraints. Harris and Kolen (1989) used bootstrap methodology to examine the stability of the Delta method across 250 resamplings. The authors found only moderate stability for the Delta method even with sample sizes of 400 in both the reference and focal groups. Similarly, in a comparison of four item bias screening statistics, Perlman et al. (1988) found the Delta method to have the lowest correlation with other statistics proposed for bias detection.

In contrast to the Delta method, the Standardization method is generally recommended for use only with relatively large samples of examinees (Dorans, 1989; Scheuneman & Bleistein, 1989). Dorans and Kulik (1986) suggested that confusing results in an item bias study may have been attributable to an "inadequate" number of focal group members, when the size of the focal group was 2,616. Hills (1989) stated that the Standardization method is normally used in testing programs with 10,000 or more examinees, and he recommended a minimum of 5,000 examinees in order to appropriately apply this technique. However, the Educational Testing Service (ETS) routinely uses the Standardization method to examine the SAT for item bias. For pre-trial items, ETS requires only 100 examinees in the focal group and 500 total examinees; for final form items, 200 examinees are required in the focal group and 600 examinees in all (Schmitt, personal communication, July 25, 1991).

The Mantel-Haenszel technique has been suggested as a more appropriate method for use with relatively small samples

(Buhr & Legg, 1990; Dorans, 1989; Holland & Thayer, 1986). Hills (1989) suggested that the Mantel-Haenszel method could be used with as few as 100 examinees in each group. DeMauro (1990) reiterated ETS' requirement of 200 focal group examinees and 600 total examinees. However, Perlman et al. (1988) found relatively low reliability of item bias identification across 30 replications for both the Mantel-Haenszel and the Delta methods, when the samples consisted of 200 to 300 examinees in each group. The reported reliability indices ranged from .38 to .58 for the Mantel-Haenszel, and from .39 to .53 for the Delta method.

Additional research is needed to assess the relative performance of item bias detection techniques applied to small samples of examinees. The information that is currently available to administrators of testing programs that service limited numbers of examinees is inadequate for making informed decisions about item bias screening techniques. To this end, the Monte Carlo study reported in this paper was undertaken. The relative effectiveness in small samples of the Delta method and the modified versions of the Mantel-Haenszel and Standardization methods described above were evaluated. The remainder of this paper describes the method, results and implications of this study.

Method

Pseudo-populations Examined

The data on which this research was conducted were random samples drawn with replacement from existing archival test data files. These files represented pseudo-populations from which samples were drawn. Data files from three teacher

subject area tests were used: Elementary Education (1-6), Early Childhood Education (K-3), and Specific Learning Disabilities (K-12). These pseudo-populations will be referred to as test forms 1, 2 and 3, respectively, in the remainder of this report. The number of multiple-choice items on the test forms were 140, 141, and 118 for forms 1, 2, and 3, respectively.

Induction of Item Bias

Within each pseudo-population, item bias was experimentally controlled using the following procedure. Each pseudo-population was randomly divided into two files, one with 80% of the examinee records (representing the reference group) and the other with 20% of the examinee records (representing the focal group). Using this random division of records provides data files in which differences in item difficulty occur only by chance. To verify the equivalence of item difficulty in the two groups, the p-values were compared prior to the induction of bias.

In each pseudo-population, nine items were selected for bias induction. Items were selected to represent all combinations of high, moderate and low values of both difficulty and discrimination (i.e., one item was selected at low values of both difficulty and discrimination, a second item was selected at a low value of difficulty and a moderate value of discrimination, etc.). Low values of item difficulty were below 0.30, moderate values were between 0.30 and 0.70, and high values were greater than 0.70. Low values of item discrimination were below 0.20, moderate values were between 0.20 and 0.35 and high values were greater than 0.35.

Bias was induced in the items by, first, stratifying the reference and focal groups on the basis of total test score and, second, modifying the item responses of randomly selected records within each stratum of each of the populations. The induced bias was designed to yield a difference in item difficulty favoring the reference group. For the randomly selected records, correct responses in the focal group were changed to incorrect, and the reverse was followed for the reference group. The stratification of the data files by total test score was used to maintain the item discrimination indices at the level obtained before the bias induction. Three levels of bias (magnitude of difference in item p-value between the reference and focal pseudo-populations) were examined in this research: 0.1, 0.2, and 0.3. The effects of each level of bias were examined in separate replications of the study. That is, in the first execution, all nine items were induced to a 0.1 level of bias; in the second, all nine items were induced to a 0.2 level; and in the third execution, all nine items were induced to a 0.3 level.

After the item bias was induced in the pseudo-population files, random samples were drawn with replacement, and the three bias detection indices were computed for each item in each sample. One thousand samples were drawn of size 50, 100, and 200. The samples were drawn such that 80% of the records in each sample were obtained from the reference group and 20% of the records in each sample were obtained from the focal group. Thus, for example, a sample size of fifty represents 40 observations from the reference group and 10 observations from the focal group.

Results

The overall descriptive statistics for the three bias indicators are presented in Table 1. For each sample size and level of imputed bias, the mean and standard deviation of each bias indicator, computed over test items and the 1000 replications, is presented. These descriptive statistics were computed separately for the bias-induced items and the non-induced items within each pseudo-population.

Insert Table 1 about here

To examine the sensitivity of each statistic to biased items, effect sizes were calculated. The effect size of each statistic is the ratio of the difference between the mean values of the statistic for bias-induced items and unbiased items to the empirical estimate of the standard error of the statistic. Thus, the effect size is given by

$$\text{Effect Size} = \frac{\hat{\mu}_b - \hat{\mu}_u}{\hat{\sigma}_u}$$

where

$\hat{\mu}_b$ = mean value of the statistic for bias-induced items,
$\hat{\mu}_u$ = mean value of the statistic for unbiased items, and
$\hat{\sigma}_u$ = standard error of the statistic for unbiased items.

The effect size for each statistic, under each of the conditions examined in the study are presented in the right column in Table 1. For all three statistics, the effect size increases as (a) sample size increases, and (b) the magnitude of the bias increases.

In comparing the effect sizes across the three statistical indices, the effect size of the Mantel-Haenszel chi-square statistic is consistently the largest, and the effect size of the Delta statistic is consistently the smallest. In samples of size 50, the best performance by the Delta method was with test form 2 and a bias effect of 0.3, which yielded an effect size of 0.70 (i.e., the mean value of the statistic for biased items was seven-tenths of a standard error higher than the mean of the statistic for the unbiased items). In contrast, for samples of size 50 from test form 2 and a bias effect of 0.3, the Standardized difficulty index yielded an effect size of 1.71, and the Mantel-Haenszel yielded an effect size of 3.68.

Across all of the conditions examined in the experiment, the best performance by the Delta method was obtained with samples of size 200 from exam form 1 with a bias effect of 0.3. In this condition the effect size of Delta was a substantial 3.29. However, under this condition the effect sizes for the Standardized difficulty index and the Mantel-Haenszel were 5.00 and 8.27, respectively.

Proportion of Correct Identification of Biased Items

For each of the 1000 samples evaluated in each condition examined (each combination of sample size, examination form, and level of bias), the nine items with the most extreme

value of each statistic were flagged as potentially biased. An intuitive index of the overall success of each bias detection statistic is the proportion of these flagged items (bias-declared items) that were the bias-induced items in the pseudo-populations. Optimal performance by a screening statistic would result in this index reaching a value of one (when the nine bias-declared items in every sample are the nine bias-induced items). The overall success rate for each screening statistic in biased item detection is presented in Table 2.

Insert Table 2 about here

The results of this analysis parallel those obtained from the examination of the effect sizes of the indices. The best performance in biased item detection was obtained from the Mantel-Haenszel and the worst performance from the Delta method. Note that in the examination of the proportion of items correctly identified as biased, the difference between the performance of the Mantel-Haenszel and the performance of the Standardized difficulty method appears to be trivial. In 20 of the 27 examination conditions examined in the experiment, the difference between the success rates was 5% or less. However, under only one of the conditions examined (test form 3, samples of size 200, and a bias effect of 0.3) the Standardization technique outperformed the Mantel-Haenszel (73% vs. 72% success). In the analysis of the effect sizes of the statistics, the difference between these two statistics was notably more pronounced.

For the Delta method, fewer than one-half of the bias-

declared items were bias-induced items in 25 of the 27 conditions examined. Only with samples of size 200 from exam forms 1 and 2 and a bias effect of 0.3 did this technique exceed 50% success. The Standardized difficulty index exceeded 50% success in only seven of the 27 conditions examined, and the Mantel-Haenszel exceeded 50% correct in nine of the 27 conditions.

To further investigate the success of each statistic at identifying biased items, the success rate of each index for each of the bias-induced test items was examined. Summary data on the overall success rates (collapsing across sample sizes) are presented in Table 3.

Insert Table 3 about here

In comparing the performance of the indicators presented in Table 3, the similarity between the Mantel-Haenszel and the Standardization methods is evident, as is the superiority of either of these methods to the Delta technique. The Delta technique shows its greatest effectiveness at identifying bias in items that have high p-values. This effect is attributable to the inverse-normal transformation of the group p-values used in this technique. This transformation accentuates differences in item difficulties at the extreme values (p-values near zero or one). For the three items with high p-values on exam form 2, the Delta method outperformed the Standardization technique consistently, as it did on one of the high p-value items on test form 1. However, the Delta method did not outperform the Mantel-Haenszel on any of the items examined.

The similarity in the performance of the Mantel-Haenszel and the Standardization method is evident in the data presented in Table 3. In 48 of the 81 conditions summarized in this table (59% of the conditions), the Mantel-Haenszel statistic was more effective in bias identification than the Standardization method. However, the difference in effectiveness is typically quite small. Of the 33 experimental conditions in which the Standardization technique outperformed the Mantel-Haenszel, 25 of the conditions were in bias detection of items of moderate difficulty.

Discussion and Recommendations

With samples of size fifty, none of the bias screening statistics examined reached an overall fifty percent success rate at bias detection. The best performance with these small samples was achieved by the Mantel-Haenszel method. With a bias level of 0.3, the bias-induced items flagged by this statistic were identified as biased items 47% of time for exam form 1, 45% of the time for exam form 2, and 37% of the time for exam form 3. The corresponding rates for the Standardization method were 44%, 33%, and 36%. The Delta method showed much lower rates of successful bias detection (18%, 19%, and 11% for the three examinations in samples of size 50 and a bias level of 0.3).

When the sample size increased to 100, both the Mantel-Haenszel and the Standardization methods exceeded 50% success at bias level 0.3. At this sample size, the Mantel-Haenszel statistic reached or exceeded 40% success for the bias level of 0.2, while the Standardization method remained between 30% and 40% success at this bias level. The performance of the Delta method was notably lower than

these levels.

Finally, at samples of size 200, the Mantel-Haenszel's success rate exceeded 70% with bias levels of 0.3, and exceeded 50% with bias levels of 0.2, but remained below 30% for a bias level of 0.1. The Standardization method's success rate also exceeded 70% with bias levels of 0.3, but dropped below 50% success with a bias level of 0.2, and dropped below 20% with a bias level of 0.1. At this sample size and a bias level of 0.3, the Delta method's success rate exceeded 50% on two of the three examinations.

Although the success rates with small samples seem low, the probability of an item being flagged as biased in this experiment (i.e., being one of the nine items with the most extreme value of the statistic), if items are responding randomly to the statistic, is just over 6% for a test with 140 items, and just under 8% for a test with 118 items. These chance rates were clearly exceeded even when only ten focal group examinees were included in the sample (samples of size 50).

In comparing the three statistical indices, the Mantel-Haenszel is the best performer. Although its performance advantage over the Standardization method is slight, the advantage is consistent over the conditions examined in this study. The only exception to this advantage is the detection of bias in items of moderate difficulty level, when the Standardization method outperformed the Mantel-Haenszel. The most striking outcome of this research is the clear superiority across all conditions examined of both the Mantel-Haenszel and the Standardization method to the Delta method. Previous research has shown that the Delta method

correlates poorly with other bias detection indices (Buhr & Legg, 1990; Perlman et al., 1988). However, because these studies did not use experimentally-controlled, induced item bias, no criterion was available for evaluating the accuracy of the various bias detection methods under investigation. This study clearly shows that not only does the Delta method correlate poorly with the Mantel-Haenszel and the Standardization methods, it is also much less accurate in the identification of biased items.

Shepard et al. (1985) point out two classes of weaknesses in much of the research which has been done on item bias indices. In the first, simulated data is used so that definitive knowledge about which items are biased is available. However, the simulated data in these studies may not accurately resemble the performance of real examinees. In the second type of study, real data are used; in these studies, however, the researcher has no prior knowledge of which items (if any) should be classified as biased. The present study has attempted to overcome both of these methodological weaknesses by retaining the advantages of using real data, while inducing item bias in order to gain the benefit of accurate information regarding the correctness of the bias indices.

In summary, previous research suggesting that item bias indices such as the Mantel-Haenszel and Standardization methods should only be applied to large samples may have been overly conservative. The results of this study support the use of statistical screening for item bias, even with samples as small as 50 examinees, and with only ten focal group members in each sample. The proportion of items correctly identified as biased by all indices examined

increased with both sample size and the magnitude of the bias. The statistical index evidencing the best performance at bias detection was the Mantel-Haenszel statistic, followed closely by the Standardization method. Either of the indices clearly outperformed the Delta method.

References

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

Buhr, D. C. & Legg, S. M. (1990). The investigation of methodology to screen for differentially functioning items within the Florida statewide testing programs. Gainesville, FL: University of Florida, Final Report to the Institute for Student Assessment and Evaluation.

DeMauro, G. E. (1990, April). Effects of representation of gender groups in the examinee population on the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2, 217-233.

Dorans, N. J. & Kulik, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Harris, D. J. & Kolen, M. J. (1989). Examining the stability of Angoff's Delta item bias statistic using the bootstrap. Educational and Psychological Measurement, 49, 81-87.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement Issues and Practice, 8, 5-11.

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. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Perlman, C. L., Bezrucsko, N., Junker, L. K., Reynolds, A. J., Rice, W. K., & Schulz, E. M. (1988). Investigating the stability of four methods of evaluating item bias. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans.

Scheuneman, J. D. & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

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

Table 1

Means, Standard Deviations, and Effect Sizes of Three Bias Detection Indices

Exam Form	Sample Size	Bias Level	Mantel-Haensel				Standardized Difficulty				Delta				Effect Sizes		
			Unbiased Items		Biased Items		Unbiased Items		Biased Items		Unbiased Items		Biased Items				
			MN	SD	MN	SD	MN	SD	MN	SD	MN	SD	MN	SD	M-H	Stand. P	Delta
1	50	0.1	1.36	1.70	2.39	3.02	0.12	0.09	0.15	0.11	1.34	1.00	1.20	0.88	0.61	0.39	-0.14
1	50	0.2	1.34	1.66	4.05	4.67	0.12	0.09	0.21	0.14	1.35	1.01	1.51	1.04	1.63	1.04	0.16
1	50	0.3	1.29	1.62	6.58	7.44	0.12	0.09	0.29	0.15	1.35	1.00	2.00	1.26	3.26	1.93	0.64
1	100	0.1	1.16	0.89	1.94	1.34	0.08	0.06	0.12	0.09	0.90	0.70	0.93	0.64	0.87	0.62	0.04
1	100	0.2	1.12	0.85	3.29	2.49	0.08	0.06	0.20	0.11	0.90	0.70	1.41	0.80	2.56	1.78	0.72
1	100	0.3	1.09	0.84	5.88	4.83	0.08	0.06	0.29	0.11	0.92	0.70	2.08	0.93	5.73	3.24	1.65
1	200	0.1	1.06	0.53	1.76	0.75	0.06	0.04	0.11	0.07	0.59	0.47	0.81	0.51	1.34	1.06	0.47
1	200	0.2	1.03	0.51	2.90	1.37	0.06	0.04	0.19	0.08	0.60	0.47	1.44	0.62	3.70	2.95	1.78
1	200	0.3	1.00	0.50	5.13	2.83	0.06	0.05	0.29	0.08	0.62	0.47	2.18	0.65	8.27	5.00	3.29
2	50	0.1	1.31	1.70	2.34	2.98	0.12	0.09	0.14	0.11	1.36	1.07	1.34	0.93	0.61	0.30	-0.01
2	50	0.2	1.26	1.61	4.06	5.58	0.12	0.09	0.19	0.13	1.37	1.08	1.62	1.05	1.74	0.83	0.24
2	50	0.3	1.22	1.58	7.03	8.72	0.12	0.09	0.27	0.15	1.37	1.07	2.13	1.25	3.68	1.71	0.70
2	100	0.1	1.12	0.91	1.90	1.42	0.08	0.06	0.11	0.08	0.95	0.73	1.02	0.65	0.87	0.44	0.09
2	100	0.2	1.08	0.88	3.37	2.71	0.08	0.06	0.18	0.10	0.96	0.73	1.45	0.75	2.59	1.46	0.67
2	100	0.3	1.04	0.83	6.12	4.71	0.09	0.06	0.26	0.11	0.96	0.73	2.06	0.89	6.09	2.76	1.51
2	200	0.1	1.02	0.54	1.70	0.77	0.06	0.05	0.09	0.06	0.67	0.50	0.84	0.49	1.28	0.70	0.34
2	200	0.2	0.99	0.52	2.95	1.45	0.06	0.05	0.17	0.08	0.68	0.51	1.42	0.58	3.75	2.35	1.47
2	200	0.3	0.96	0.51	5.34	3.02	0.06	0.05	0.26	0.08	0.69	0.51	2.13	0.64	8.52	4.28	2.83
3	50	0.1	1.30	1.86	2.28	2.69	0.11	0.08	0.14	0.10	1.56	1.20	1.41	0.99	0.53	0.43	-0.12
3	50	0.2	1.27	1.80	3.52	4.08	0.11	0.08	0.18	0.12	1.56	1.20	1.57	1.05	1.25	0.92	0.01
3	50	0.3	1.22	1.75	5.39	6.47	0.11	0.08	0.24	0.13	1.56	1.20	1.90	1.16	2.39	1.65	0.29
3	100	0.1	1.13	1.03	2.10	1.57	0.08	0.06	0.11	0.08	1.10	0.83	0.98	0.71	0.94	0.62	-0.14
3	100	0.2	1.08	0.95	3.33	2.43	0.08	0.06	0.17	0.09	1.11	0.83	1.28	0.83	2.36	1.51	0.20
3	100	0.3	1.06	0.59	5.28	0.51	0.08	0.06	0.24	0.10	1.13	0.84	1.74	0.97	7.14	2.68	0.73
3	200	0.1	1.04	0.63	1.92	1.04	0.06	0.04	0.10	0.06	0.82	0.58	0.78	0.49	1.41	0.88	-0.07
3	200	0.2	0.99	0.60	3.30	1.96	0.06	0.04	0.16	0.07	0.82	0.58	1.17	0.61	3.88	2.37	0.60
3	200	0.3	0.97	0.59	5.53	3.07	0.06	0.04	0.24	0.07	0.85	0.58	1.73	0.73	7.78	4.02	1.52

Table 2

Proportion of Biased Items Correctly Identified by Three Bias Indices

Exam Form	Sample Size	Bias Level	Proportion of Correctly Classified Biased Items		
			M-H	Stand. P	Delta
1	50	0.1	0.16	0.13	0.04
1	50	0.2	0.30	0.26	0.09
1	50	0.3	0.47	0.44	0.18
1	100	0.1	0.20	0.18	0.06
1	100	0.2	0.45	0.40	0.18
1	100	0.3	0.69	0.65	0.34
1	200	0.1	0.29	0.26	0.15
1	200	0.2	0.63	0.60	0.41
1	200	0.3	0.87	0.86	0.65
2	50	0.1	0.15	0.12	0.05
2	50	0.2	0.29	0.21	0.10
2	50	0.3	0.45	0.38	0.19
2	100	0.1	0.19	0.14	0.07
2	100	0.2	0.42	0.33	0.18
2	100	0.3	0.64	0.57	0.32
2	200	0.1	0.26	0.19	0.13
2	200	0.2	0.57	0.49	0.32
2	200	0.3	0.79	0.76	0.51
3	50	0.1	0.15	0.14	0.04
3	50	0.2	0.26	0.23	0.07
3	50	0.3	0.37	0.36	0.11
3	100	0.1	0.21	0.18	0.05
3	100	0.2	0.40	0.34	0.12
3	100	0.3	0.55	0.54	0.21
3	200	0.1	0.27	0.22	0.06
3	200	0.2	0.53	0.48	0.17
3	200	0.3	0.72	0.73	0.31

Table 3
Proportion of Items Correctly Identified As Biased By Three Statistical Indices
for Each Level of Item Discrimination and Difficulty

Examform= 1									
		Bias=0.1			Bias=0.2			Bias=0.3	
		Item Bias Index			Item Bias Index			Item Bias Index	
Item Discrimination	Item Difficulty	M-H	Stan. P	Delta	M-H	Stan. P	Delta	M-H	Stan. P	Delta
Low	Low	0.225	0.239	0.043	0.469	0.503	0.158	0.691	0.739	0.327
Low	Middle	0.204	0.265	0.055	0.432	0.516	0.144	0.617	0.709	0.276
Low	High	0.280	0.146	0.201	0.587	0.408	0.474	0.794	0.666	0.698
Middle	Low	0.169	0.166	0.029	0.398	0.370	0.098	0.644	0.609	0.257
Middle	Middle	0.172	0.206	0.038	0.350	0.398	0.112	0.544	0.614	0.207
Middle	High	0.260	0.175	0.152	0.531	0.423	0.386	0.730	0.642	0.598
High	Low	0.165	0.145	0.022	0.404	0.344	0.100	0.661	0.592	0.238
High	Middle	0.202	0.204	0.052	0.419	0.413	0.165	0.622	0.618	0.295
High	High	0.282	0.161	0.157	0.550	0.406	0.383	0.754	0.645	0.623

Examform= 2									
		Bias=0.1			Bias=0.2			Bias=0.3	
		Item Bias Index			Item Bias Index			Item Bias Index	
Item Discrimination	Item Difficulty	M-H	Stan. P	Delta	M-H	Stan. P	Delta	M-H	Stan. P	Delta
Low	Low	0.118	0.121	0.017	0.308	0.289	0.039	0.546	0.551	0.117
Low	Middle	0.230	0.266	0.060	0.450	0.501	0.152	0.655	0.712	0.296
Low	High	0.418	0.189	0.299	0.724	0.486	0.596	0.866	0.710	0.778
Middle	Low	0.279	0.261	0.029	0.520	0.485	0.093	0.724	0.720	0.281
Middle	Middle	0.069	0.088	0.009	0.204	0.227	0.034	0.405	0.435	0.098
Middle	High	0.177	0.060	0.090	0.492	0.248	0.330	0.738	0.516	0.594
High	Low	0.096	0.099	0.006	0.276	0.235	0.019	0.504	0.438	0.060
High	Middle	0.065	0.083	0.008	0.209	0.214	0.031	0.398	0.406	0.100
High	High	0.381	0.150	0.226	0.687	0.392	0.506	0.834	0.638	0.736

Examform= 3									
		Bias=0.1			Bias=0.2			Bias=0.3	
		Item Bias Index			Item Bias Index			Item Bias Index	
Item Discrimination	Item Difficulty	M-H	Stan. P	Delta	M-H	Stan. P	Delta	M-H	Stan. P	Delta
Low	Low	0.239	0.071	0.041	0.397	0.196	0.104	0.405	0.339	0.182
Low	Middle	0.145	0.244	0.006	0.257	0.102	0.010	0.445	0.592	0.027
Low	High	0.447	0.279	0.257	0.692	0.531	0.491	0.851	0.719	0.673
Middle	Low	0.161	0.148	0.012	0.355	0.313	0.029	0.584	0.572	0.091
Middle	Middle	0.169	0.260	0.005	0.322	0.443	0.022	0.500	0.641	0.041
Middle	High	0.151	0.078	0.061	0.389	0.252	0.148	0.612	0.464	0.338
High	Low	0.371	0.300	0.039	0.561	0.464	0.092	0.565	0.641	0.203
High	Middle	0.078	0.155	0.004	0.186	0.275	0.004	0.345	0.467	0.009
High	High	0.143	0.078	0.049	0.409	0.255	0.148	0.617	0.452	0.319

Author's Notes

The research reported herein was supported in part by a grant from the Florida Department of Education (DOE) and the Institute for Instructional Research and Practice (IIRP) at the University of South Florida. However, the opinions expressed are those of the authors and do not reflect the position or policy of the DOE or IIRP.