DOCUMENT RESUME

ED 287 882 TM 870 639

AUTHOR Lautenschlager, Gary J.; Park, Dong-Gun
TITLE A Simulation Study of the Effects of Ability Range Restriction on IRT Item Bias Detection Procedures.
PUB DATE Apr 87
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; *Item Analysis; *Latent Trait Theory; Mathematical Models; Sample Size; Statistical Bias; *Test Bias; Test Length
IDENTIFIERS *Range Restriction

ABSTRACT

The effects of variations in degree of range restriction and different subgroup sample sizes on the validity of several item bias detection procedures based on Item Response Theory (IRT) were investigated in a simulation study. The degree of range restriction for each of two subpopulations was varied by cutting the specified subpopulation ability distribution at different locations and retaining the upper portion of the distribution. It was found that range restriction did have an effect on the accuracy of the bias detection procedures. The signed area index was least influenced by variations in range restriction, whereas the base low area index was found to be invalid regardless of the degree of range restriction. The findings for variations in sample size were mixed. (Author)

A Simulation Study of the Effects of Ability Range Restriction On IRT Item Bias Detection Procedures

Gary J. Lautenschlager and Dong-Gun Park

University of Georgia

Address correspondence to:

Gary J. Lautenschlager

Department of Psychology

University of Georgia

Athens, GA 30602

(404) 542-3054

ABSTRACT

The effects of variations in degree of range restriction and different subgroup sample sizes on the validity of several item bias detection procedures based on Item Response Theory (IRT) were investigated in a simulation study. The degree of range restriction for each of two subpopulations was varied by cutting the specified subpopulation ability distribution at different locations and retaining the upper portion of the distribution. It was found that range restriction did have an effect on the accuracy of the bias detection procedures. The signed area index was least influenced by variations in range restriction, whereas the base low area index was found to be invalid regardless of the degree of range restriction. The findings for variations in sample size were mixed.

A Simulation Study of the Effects of Ability Range Restriction On IRT Item Bias Detection Procedures

Numerous statistical techniques have been proposed for detecting test item bias (Green and Draper, 1972; Angoff and Ford, 1973; Wright et al., 1976; Rudner, 1977; Wright and Stone, 1979; Scheuneman, 1979; Camilli, 1980; Lord, 1980; Linn, Levine, Hastings, and Wardrop, 1981; Hulin et al., 1982). Many of these techniques pinpoint bias at the item level. The nature of these techniques has led some psychometricians to refer to them as item bias detection methods. Some methods are adapted from classical test theory (CTT) and others from item response theory (IRT). CTT based techniques are theoretically problematic, as most of th m have relied on sample dependent item statistics from CTT. The most frequently used CTT index, the transformed item difficulty approach, uses the CTT item difficulty index, p_i. Although previous investigators have attempted to control for the sample dependence property of the statistic, this has not proven entirely successful (Ironson, 1982). However, the techniques based on item response theory are said to be theoretically superior to CTT ones. The sample free quality of IRT item parameters tend to make those methods of item bias detection less sensitive to distributional differences in subpopulation samples (Lord, 1980; Shepard, Camilli & Averill, 1981; Shepard, Camilli & Williams, 1984).

All of the item bias techniques developed to date, whether IRT based or CTT based, are dependent on information internal to the test in question. These techniques establish the standard of unbiasedness by using either the total score on the test or the estimated trait score based on the responses to the items on the test. Bias in one item cannot be identified without considering information from other items on the

test. Bias is identified when an item exhibits different characteristics than the rest of the items in the test (Shepard, 1982; Burrill, 1982). The appropriateness of an item to the trait assumed to be measured by the test is not addressed by these techniques (Petersen, 1977; Shepard, 1982; Shepard, et al., 1985). The techniques only serve to make a test homogeneous, whether or not the test thus constructed measures what we want it to measure. Many authors have suggested that both judgmental and empirical methods must be used in combination.

An IRT model provides a probabilistic way of linking individuals' item responses to the latent characteristic assumed to underlie those responses. IRT models make use of an item characteristic curve which depicts the relationship between the probability of a correct response to the item and the latent characteristic. For several reasons the logistic IRT models are preferred in most practical applications of IRT. The three-parameter logistic model is mathematically given by:

$$(1) \quad P_i(\theta) = c_i + (1 - c_i) / \{ 1 + \exp[-1.7a_i (\theta - b_i)]\}$$

where $P_i(\theta)$ is the probability of an examinee with a given level of ability on the latent trait θ answering item i correctly, b_i is the difficulty of item i, a_i is the discrimination index for item i, c_i is the lower asymptote for item i and serves as a baseline for guessing.

The above equation states that probability of success on an item depends on nothing but three item parameters and examinee ability θ. If the model holds true, a person's trait θ is all we need in order to determine his/her probability of success on any given item (Lord, 1980). In other words, individuals who have the identical value on the trait dimension must have an equal probability of getting a specified item correct, regardless of their subpopulation group membership.

According to the model, therefore, the item parameters remain the same regardless of the distribution of the trait in the subpopulation tested. This notion of parameter invariance is a basis of all IRT bias detection methods, and a finding of parameter variance would mean bias exists. The invariance property can be understood by interpreting ICCs as nonlinear regression lines (Hulin et al., 1983; Lord, 1980). Note that although IRT item parameters are invariant across subpopulations, item parameter estimates will not be necessarily identical when calculated in different samples. Since the choice of origin for the trait scale is purely arbitrary, the invariance of item parameters holds true only if the origin and unit of the trait scale is the same. The estimates can be placed on the same scale via an appropriate linear equating transformation.

Factors Affecting IRT Parameterization

Validity studies are often problematic due to many artifactual factors, such as small sample size, restriction of range in the sample, and criterion unreliability. These problems may be also exist when item bias research is performed in various decision-making contexts. Ideally, item bias studies can be incorporated at the test construction stage to minimize the chances of bias accusation arising later (Berk, 1982; Drasgow, 1982). Consequently, it is important to understand what factors can potentially affect the validity of the IRT bias detection techniques.

Other researchers have shown that CTT-based chi-square methods are sensitive to cutoffs for ability intervals, sample sizes, and the distribution of the total test scores. Nungster (1977) and Rudner (1977) demonstrated that the chi-square values can become quite inflated when the total observed score distributions differ. Baker (1981) also noted that

the Scheueman chi-square procedure is confounded by unequal sample sizes for the two subpopulations.

There are few Monte Carlo or empirical studies of this kind for IRT bias techniques. It is clear that sample size and test length are important to some extent because they jointly influence the stability of estimates of person and item parameters (Hulin et al., 1983). Furthermore, it is not unreasonable to expect that true ability distributions, like observed score distributions, may be highly skewed if not truncated in numerous applied settings.

As noted previously, parameter estimates in IRT remain the same across subpopulations within a linear transformation. It should not be assumed, however, that parameters can be estimated with equal accuracy on all the subpopulation samples of interest (Ironson, 1982). As a matter of fact, estimation of person and item parameters are an important problem encountered in applications of IRT models. Since their true values are unknown, they must be estimated simultaneously. LOGIST uses an iterative procedure alternating between trait estimates and item parameter estimates. The iteration is continued until both item and trait paramter estimates converge, that is, until the estimates change by an arbitrarily small value from the i^{th} to $i^{th} + 1$ iteration (Hulin et al., 1983).

Wright (1977) argued that irresovable problems arise for some IRT models when item and trait parameters must be estimated simultaneously (Hulin et al, 1983). However, Hulin et al.(1982) showed that simultaneous estimation of trait and item parameters using an iterative procedure such as LOGIST may be sufficiently accurate for many practical applications of IRT. This sufficient accuracy, however, is obtained within constraints of sample size and number of items. Although specific

sample size and test length depend on which parameters one wishes to estimate accurately and/or the purposes for which IRT is used for, Swaminathan and Gifford (1980) and Lord (1980) recommended using as many as 50 items and 1,000 persons for the recovery of item parameters. However, Hulin, Lissak, and Drasgow (1982) have investigated the recovery of an ICC, finding that a test as short as 30 items with a sample size of 1,000 examinees or 60 items and 500 examinees for the three-parameter data, appear sufficient for accurate estimation of ICCs.

Criterion-related validity studies are limited by range restriction, because they need external criterion scores as well as predictor test scores. Since item bias studies do not need external criterion scores, it is expected that the situations where the effect of the range restrictions occur in the validity studies are not necessarily parallel to the ones in IRT item bias research. It is not unlikely that range restriction on ability, direct or indirect, would occur in many practical situations.

The purpose of the present study is to investigate the effect of range restriction in the subpopulation groups on the validity of item bias detection procedures. A type of range restriction similar to that which might be encountered in selection/training contexts provided the focus for the study. The restricted ranges of traits for each subpopulation were obtained by arbitrarily cutting the specified population ability distributions and taking the upper part of the distributions.

The sample size in the subpopulation comparison groups was also manipulated. Although sample size and test length should be considered together for accurate estimation of IRT parameters, requirements for test length are not as severe when compared to the sample size requirements. Hulin et al. (1982) indicated that for research involving the comparison

of ICCs, such as item bias studies, large numbers of items are not needed. However, large numbers of subjects are necessary. Therefore, in the present investigation test length was fixed at 50 items and only sample size and range restriction were varied. Hulin et al. (1982) reported that there were tradeoffs between test length and sample size. This might also be true of sample size and degree of range restriction.

The second question investigated in the study concerned estimation accuracy of the item response function (IRF). All of the IRT based bias detection indices compare IRFs calculated based on item parameters generated separately from each relevant subpopulation. Accuracy of these techniques are very closely related to estimation accuracy of the IRFs. In effect, examining the IRF estimation accuracy for the subpopulations would be a desirable step before determining the efficiency of the techniques. In the present stduy examination of the IRFS was important in its own right, providing an indication of how restriction of range in traits affected the IRF estimation accuracy.

In summary, the following questions were addressed in the present study:

1. To what extent did restriction of range in subpopulation samples affect the accuracy of the selected bias detection indices? Did different combinations of range restrictions for the two subpopulation samples produce distinct results?

2. When range restriction and sample size were considered jointly, what were the effects in terms of accuracy of parameterization?

3. How accurately were the IRFs estimated?

METHOD

Data Simulation

Binary response data sets were generated using a modified (i.e., the c-parameter was fixed) three-parameter logistic IRT model. The generation of responses began by constructing biased items. In order to make items biased, item parameters were manipulated to be subpopulation dependent. Two 50 item four-option multiple choice ability tests are simulated with differential overall test bias. Test 1 was constructed to contain 40 biased items and Test 2 had 10 biased items.

For both tests the item parameters for subpopulation 1, a_{i1} and b_{i1}, were drawn from uniform distributions in the interval [+.65, +1.6] and the interval [-3,+3], respectively. The values of a_{i2} and b_{i2} for subpopulation 2 are created by subtracting the randomly sampled values of $(a_{i1} - a_{i2})$ and $(b_{i1} - b_{i2})$ from the item paramters for subpopulation 1. In generating random values for $(a_{i1} - a_{i2})$, the values were constrained to positive numbers. To avoid difficulty in estimating c parameters, these were arbitrarily fixed at .20 for all items in both subpopulations.

The distribution of the latent trait was assumed to be normal with a standard deviation of 1.0 for each subpopulation. Subpopulation 1, hereafter denoted as the 'A' group had a mean ability of +.5 and subpopulation 2, hereafter denoted as the 'B' group, had a mean ability of -.5. Test length was fixed at 50 items, because in practice length of test is less a problem than sample size. Within each subpopulation four restriction groups (labeled W, X, Y and Z) and a no resriction group (labeled N) of examinees were specified as follows: (W) all examinees with θ above +1.5 (X) all those with θ above +1.0 (Y) all those with θ above 0.0 (Z) all those with θ above -1.0 (N) no restriction.

The probability of a correct response to an item was calculated by entering appropriate subpopulation item parameters and the trait values of simulated examinees into equation 1. The probability of a correct response was then compared to a random number drawn from the (0, 1.0) uniform distribution. If the sampled number was larger than the probability of a correct response, then the item was scored as incorrect: otherwise, a correct response was specified for the item. The item parameters a_i, b_i, and the person parameter θ were estimated by analyzing the simulated data sets using the LOGIST (Wingersky, Barton & Lord, 1982) computer program. Default convergence criteria for LOGIST were used.

The known simulated amount of bias in an item was measured by each of the bias detection indices described earlier, and this value was correlated with detected amount of bias measured by the same index in the simulated item responses. Preliminary analysis showed that when the correlations were calculated using all 50 items, relatively low correlations were observed for most of the indices in the study, and the correlations were also less lawfully behaved with changes in degree of restriction. It was suspected that poor estimates of item parameters had hindered a clear effect of range restriction. Previous research by others (Swaminathan & Gifford, 1983; Hulin et al., 1982) has indicated that the range of b-parameter values in test items was an important factor that determined the accurate estimation of item parameters. Therefore, in addition to examining the correlations for all 50 items, the correlations for a subset of 27 items that were selected from the original 50 items, and whose difficulty parameter values were between -2.0 and +2.0, and hence were likely to be better estimated, were also computed.

Within each restricted subpopulation, three sample sizes of N = 1000, 600 and 300 examinees were sampled. Each restriction group-sample size

combination for group 'A' was paired with the restriction-sample size combinations for group 'B' which had an identical restriction conditions with equal or smaller sample sizes. This made group 'A' a majority group. Applying these criteria, a total of 60 comparison datasets were generated to permit 30 comparisons of an 'A' group dataset with a 'B' group dataset.

Equating

Since the trait scales from an IRT analysis are arbitrary, the a and b values for the two subpopulations are not directly comparable. ICCs must first be equated to the same scale. To make the adjustment of item parameters from different subpopulations, a linear transformation of the b parameters as described in Linn, et al. (1981) was used. The equating is determined by a best fitting line that adjusts for the difference in average in b-parameter values and has a slope equal to the ratio of the standard deviations of the two sets of b's (Shepard et al., 1984). Linn et al. (1981) selected equating constants so that the weighted mean and variance of the b's in the comparison group were equal to the weighted mean and variance of the b's in the base group. The weight for each item was determined by taking the inverse of the larger sampling variance for the b-parameter from either the base or comparison group.

Estimation Accuracy of Item Response Function

As noted previously, all of IRT based bias detection indices compare IRFs calculated from the relevant subpopulations. Accuracy of these techniques are very closely related to the estimation accuracy of IRFs. In effect, examining the IRF estimation accuracy for the subpopulation groups would be a desirable step before determining efficiency of the techniques. If IRFs proved poorly estimated, but bias detection was

accurate, the accuracy might have been obtained by chance or the efficiency of the techniques might be questionable. This examination would be also important in its own right, giving an indication of how restriction of range on a trait affected the IRF estimation accuracy.

As noted by Hulin et al. (1983), the primary emphasis in a study of item bias is on the accuracy of the estimation of ICCs. Although a close relation between parameters and parameter estimates can indicate good estimation, this criterion may be overly stringent for a study of bias. They suggested the statistic, the Root Mean Square Error (RMSE) for use in investigating the recovery of ICCs rather than the recovery of item parameters. In the present study this statistic was used as an index of IRF estimation accuracy. The index is expressed in the following equation as:

$$RMSE = \{ 1 / 31 \sum_{j=1}^{31} [P_i(\theta_j) - P_i(\theta_j)]^2 \}^{1/2}$$

where $P_i(\theta)$ is the true ICC for item i; $P_i(\theta)$ is the recovered ICC. Thirty-one values for θ were chosen at equal intervals from -3 to +3. RMSEs were averaged over all 50 items in each range restriction sample size combination.

Bias Indices

Hambleton and Swaminathan (1985) divided procedures for assessing item bias using IRT methods into three categories involving, comparisons of ICCs, comparisons of vectors of item parameters, and comparisons of the fit of the IRT model to the data. Since the data are simulated, the RMSEs were used to address the last concern. To address the other two general procedures, the following item bias indices were computed:

1. Area Indices

The Ability continuum was divided from -3 to +3 into 600 intervals (Linn et al., 1981). The absolute value of the differences of the ICCs given by:

$$D_j = \left| P_{i1}(\theta_j) - P_{i2}(\theta_j) \right|$$

at the midpoints of these 600 intervals are multiplied by the width of the interval, .01 in this case.

a.) Base-High Area (BH): $BH = \sum_{j=1} (.01)*k*D_j$

b.) Base-Low Area (BL): $BL = \sum_{j=1} (.01)*(1-k)*D_j$

where $k = 1$ if $P_{j1} > P_{j2}$, and $k = 0$ if $P_{j1} < P_{j2}$

c.) Total Area (TA): TA= BH + BL

d.) Signed Area (SA): SA= BH - BL

e.) Root Mean Squared Difference (RMSD):

$$RMSD = \left(1/600 \sum_{j=1} [P_{i1}(\theta_j) - P_{i2}(\theta_j)]^2 \right)^{1/2}$$

2. Differential Parameter Indices

Lord's chi-squared statistic was calculated for testing the null hypothesis that for a particular item i both $b_{i1} = b_{i2}$ and $a_{i1} = a_{i2}$. The chi-squared statistic is

$$\chi^2 = \underline{v}_i' \, (\underline{S}_{I1} + \underline{S}_{I2}) \, \underline{v}_i$$

where $\underline{v}_i' = (b_{i1} - b_{i2}, a_{i1} - a_{i2})$; $\underline{S}_{i1}$ is sampling variance-covariance matrix of a_{i1} and b_{i1} in group 1, and similarly for $\underline{S}_{i2}$. $\underline{S}_{i1}$ and $\underline{S}_{i2}$ are found for maximum likelihood estimators from the formulas $\underline{S}_{i1} = \underline{I}_{i1}$ and $\underline{S}_{i2} = \underline{I}_{i2}$, where $\underline{I}_1$ is the 2 x 2 information matrix for a_1 and b_i. The significance test was carried out separately for each item by computing and comparing the chi-squared value to the critical value with 2 degrees of freedom (Lord, 1980).

RESULTS

Identification of Bias

The six item bias indices were calculated for each item in each of the 60 comparison conditions created by employing the 5 levels of range restriction, 6 levels of group sample size combinations, and 2 levels of the number of biased items. The amount of "true" simulated statistical bias was calculated using each of the bias detection indices in the study and was in turn correlated with the detected bias measured by the same bias index for each sample condition. One exception was that the bias estimates computed using the chi-square index were correlated with the "true" bias measured by the total area index.[1]

The results from these analyses are presented in Tables 1 through 6, one table for each of the six bias indices used. The results are presented for a test with 10 biased items and for a test with 40 biased items. Within each test results are presented for all 50 items, and separately for the 27 items that had population b-parameter values that ranged between -2.0 and +2.0 (which presumably would have had led to better item parameter estimates in the samples). The restriction conditions and the sample sizes for the A and B groups are given in the left column in each table. Three blocks of results are presented as separate rows for which the A group sample sizes are fixed and paired with variations in the B group sample size for each possible range restriction condition, as discussed above.

Total Area Method

Test with 10 biased items

Table 1 presents correlations between the generated amounts of bias and detected amounts of bias, which were derived from the total area index

for the test that had 10 biased items. The results for equal and unequal sample sizes are discussed in separate sections below.

Equal sample size for both groups

For both 50 item and 27 item test correlations between the generated and detected amount of bias steadily decreased with an increase in range restriction in ability distribution of the samples with N=1000. This is also true for the condition with sample sizes of N=600 for both groups. The correlations between true bias and bias estimates for samples of N= 300 examinees fluctuated with an increase in restriction, but still indicated that there was an effect of range restriction on the detection accuracy of the index.

Unequal sample sizes

The results for 50-item test indicated that the index was not influenced by range restriction in the samples across different combinations of sample sizes. However, the correlations computed on 27 items whose estimated difficulties ranged beween -2.0 and + 2.0 rose with a decrease in range restriction, especially for the 1000-600 combination. The 2 restriction condition correlations were comparable to those of no restriction conditions. Interestingly enough, the W restriction condition resulted in high correlations across sample size combinations.

Test with 40 biased items

Equal sample sizes for both groups

The right-most columns of Table 1 present the results from the 40 biased item test conditions. It was expected that since biased items contribute to poorer estimation of abilities it might be expected that correlations for the item bias measures would be low. However, this was not what was found. The correlations tended to be only slightly lower in

general than in the test with 10 biased items. It was observed that as range restriction became more severe, correlations between the generated and detected amount of bias decreased when both samples had N=1000. In the equal sample size conditions where N=600 and N=300 the relationships between range restriction in the sample of examinees and the efficiency of detecting item bias did not yield a consistent pattern. Correlations for no restriction conditions were extremely low. This indicates that under the no range restriction conditions the total area index can actually be less accurate. Using a sample size of 600 for each of the two comparison groups, there was no clear indication that range restriction had a systematic effect on the index in terms of accuracy. When N=300 and the sample were restricted to theta above 1.0, the detection techniques were most effective. It is hypothesized that this result was due to large overall bias in the test. The correlations did not improve for the 27-item version of the test. This also may result from the severe distortion of the ability dimension due to too many biased items. When sample sizes equaled 300, little was learned about the effect of range restriction on the index.

Unequal sample sizes

There was little evidence that the Total Area Method is sensitive to range restriction of ability distributions, when groups are of unequal sample sizes. However, correlations were the lowest when no restriction was imposed on the ability distributions. In the extreme condition (W), relatively high correlations were observed. This was true with the both 50-item and 27-item tests.

Base-High Area Method

Test with 10 biased items

Equal sample sizes for both groups

In Table 2 it can be seen that for equal sample sizes, when N=1000 there was a substantial difference in the values of correlations between the conditions W, X and the conditions Y, Z and N. Correlations for the 50-item and 27-item tests were somewhat stable in the Y, Z, N conditions, indicating that the base-high area index was somewhat robust to restricted range in abilities. When sample sizes were equal to N=600, the trend in change was similar to that in sample sizes of N=1000 for the 50-item test. However, in the 27-item test with N=600, the index was extremely accurate in detecting item bias with a correlation of 0.944 under no restriction condition. Although relationships between the bias estimates and known bias were quite high and stable across restriction conditions, the difference in the magnitude of correlations between the groups in the presence and absence of restrictions was large. With 300 examinees in each of the groups, when there were range restrictions more severe than condition Z in the samples, the index appears to be invalid. Z proved to be a very tolerable condition in estimating degree of bias in items by the index.

Unequal sample sizes

The effects of having restricted samples for the two comparison groups seems to be clear for both 50-item test and 27-item test. As was the case with Total Area Method however, the correlations for the 50-item test were relatively low even in the absence of range restriction in samples. Therefore, the index may not always have high validity with respect to discovering the presence of bias in items. With 27 items whose simulated values of difficulty parameters lie between -2.0 and +2.0, there

was a very clear tendency of the magnitude of detectablity to become larger as the samples became less restricted in range. Interestingly, the correlations in the Z restriction conditions were the highest and higher than those in the no restriction conditions in both the 1000-300 and 600-300 sample size combinations. This suggests that not having individuals of very low ability in a sample may enhance detection of biased items.

Test with 40 biased items

Equal sample sizes for both groups

Since biased items contribute to the estimation of person parameters, it was expected that the degree of range restriction would be of little systematic influence on the bias detection accuracy of the index in the samples (see Table 4). This is not necessarily true. The outcome was similar to that obtained in the test which had 10-biased items in terms of magnitudes of correlations associated with each restriction condition. However, for the samples of N=1000 the correlations obtained in no restriction conditions were much lower than those of Z conditions. In N =300 for both groups there was a constant decrease in the detectability of the index with an increase in the degree of restriction of range in the sample.

Unequal sample sizes

There seemed to be a similar trend of the range restriction effect on the accuracy of the base-high index for unequal sample sizes. A substantial increase in the values of correlations was observed between the X and Y range restriction conditions.

Base-Low Area Method

Test with 10 biased items

Equal sample sizes for both groups

The Base-Low method resulted in near zero correlations for both 50-item and 27-item tests (see Table 3). Although the index had one correlation of 0.45 in N=600, when a Spearman correlation was computed for this condition the obtained value was zero, and thus this result may be attributed to chance.

Unequal sample sizes

The results were essentially the same as for the equal sample size conditions. All of the correlations were essentially zero, with no indication of range restriction effects.

Test with 40 biased items

Equal sample sizes for both groups

It was expected that the test with 40 biased items would have a similar outcome to the 10 biased item test, since biased items will distort the ability dimension and the index appeared to be very invalid for these particular data sets. Again, surprisingly, this was not the case. The resulting correlations were much higher than those in the test with 10 biased items, and there was an indication that correlations in the more restricted condition were lower than in the less restricted conditions. As was the case with Base-High method, the largest change in the value of the correlations occurred between X and Y conditions when N=1000 for both groups. The correlations were lower in the 27-item test than in the 50-item test, and the correlations obtained in N=600 and N=300 were close to zero.

Unequal sample sizes

Results obtained on 50-item and 27-item tests indicated no systematic influence of the restriction in range of examinee's traits. This was not true in the 1000-300 sample size combination for the 27-item test; the correlations show an increment with a decrease in the degree of range restriction in samples.

Signed Area Index

Test with 10 biased items

Equal sample sizes for both groups

The results presented in Table 4 reveal that most correlations were relatively high. However, clear increasing trends in the magnitude of the relationship between the estimated bias and known bias were not present as degree of the restriction increased. Instead, correlations remained relatively stable across restriction conditions. In contrast with the previous item bias indices discussed above, the correlations in the no restriction conditions tended to be fairly high. Focusing on the subset of 27 items, the correlations consistently increased over those obtained when all 50 items were used. All of the correlations are impressively high across the restriction conditions, indicating that this method is robust to the range restriction. One exception is when the sample sizes in both groups A and B are 300, then increasing range restriction tends to lead to a decline in the correlation.

Unequal sample sizes

As in the equal sample size condition, magnitudes of the correlations were generally the highest when examinees were sampled from a larger range of the ability distribution. The detection accuracy was quite high even in the most severe restriction case for the 27-item version of the test.

Test with 40 biased items

Equal sample sizes for both groups

Across the restriction conditions, the magnitude of the correlations were somewhat uniform and high, as was true with the test having 10 biased items (see Table 4). The correlations were on the average higher with 40 biased items in the test than with 10 biased items in the test.

Unequal sample sizes

Differences in the degree of the restriction did not appear to influence the efficiency of the method with respect to accuracy. The results were very similar for the 50-item and 27-item tests, and again the correlations were somewhat higher for the 27-item version. When N=1000 for the A groups and N=300 for the B groups, the method detected bias in items under the W restriction condition as accurately as under the no restriction condition. In summary, it may be concluded that Signed Area Method was robust to the number of biased items in the test, differences in sample sizes and the degree of range restriction in the samples.

RMSD Method

Test with 10 biased items

Equal sample sizes for both groups

Based on the results presented in Table 5, restriction in range had no consistent systematic impact on the RMSD item bias technique when considering all 50 items. However, the 27-item version yielded results more consistent with the expectation that increases in range restriction would lead to lower correlations. The correlations dropped sharply when the more severe restrictions were imposed on the range of the samples, and was most clear when N=300 in both groups.

for both groups A and B there was a large difference in the values of the

correlations in comparing the no restriction condition with the moderate restriction conditions Y and Z. There was fluctuation in the magnitude of the relationship between the generated and detected amount of bias among the restriction conditions.

Unequal sample sizes

There was no clear indication that the detection accuracy of the RSMD bias index was systematically influenced by restriction in range. The correlations under the W restriction condition were as high as under the Z or the N (no restriction) conditions. However, it cannot be concluded that the method was robust to severe restriction condition because correlations were lower under X and Y restriction conditions.

Test with 40 biased items

Equal sample sizes for both groups

There was no indication that when the distribution of the examinees of the two group samples was more restricted in range, the technique failed in discovering bias in items for 50 item test with N=1000 and N= 600 (see Table 10). With sample sizes of N=300, differences in correlations between the restriction and no restriction conditions were substantial. Across different sample size conditions, the index was the least accurate method of detecting bias in items with both the 50-item test and the 27-item test in the absence of range restriction. Except in the no restriction condition, the correlations for the 27-item test exhibited a systematic decrement as the restriction increased for any sample size combination.

Unequal sample sizes

As was the case with the identical sample sizes condition, the range restriction in samples did not seem to affect the detection accuracy of

the index. The correlations in the severest restriction condition, W, were much higher than those in the no restriction condition. Perhaps the number of biased items in the test would render results uninterpretable. In general, the RMSD index performed differently when there were 10 biased items in the test than when there were 40 biased items in the test.

Chi-Square Index

Test with 10 biased items

Equal sample sizes for both groups

With 27 items that have a value of difficulty parameter between -2 and +2, the Pearson correlations between the estimated and simulated bias showed a steady decrement as the degree of the range restriction on the ability distributions grew larger (see Table 6). The highest correlation value was obtained with 27 items under no restriction condition when the number of the examinees were 600 for the two groups. The restriction was also observed to affect the index for 50 item test in terms of efficiency.

Unequal sample sizes

With the exception of very high correlations under W conditions and 600-300 for 50 items and 1000-300 for 27 items combinations, the larger the range restriction of the sample, the more the bias detection accuracy of the technique deteriorated. Under Z conditions the method was as effective in terms of detecting bias as under N conditions.

Test with 40 biased items

Equal sample sizes for both groups

Compared to the 10 biased item test, the impact of the range restriction on the Chi Square method was clear, as the correlations fluctuated with change in the degree of the restriction. The correlations under the no restriction condition were smaller than under Z condition in

all sample size combinations. This may be attributed either to unstable estimation of abilities due to many biased items, or to poorer estimations of the item or person parameters below an ability value of -1.5. Therefore, the absence of restriction may not always be the best condition to detect item bias.

Unequal sample sizes

With the 1000-300 and 600-300 sample size combinations, decreasing the degree of the restriction led to an increase in the correlation between the generated bias and detected bias for the 50 item test. This relationship did not occur with the 27-item test. Since all of the correlations were very low when two groups differed in regard to the size of the sample, it seems that absence or presence of the range restriction in the groups does not make a practical difference in the validity of the Chi-Square index.

Root Mean Square Error

Recovered ICCs were compared to actual ICCs calculated from the simulated parameters at 31 theta values chosen at equal intervals from -3.0 to +3.0 by the root mean squared error (RMSE) for each item. RMSEs for 50 items in the test for each sample size-group combination were averaged. The results are presented in Table 7. Within each sample size, the average RMSEs for any group increased as the range restriction in the sample increased as had been expected. For the "A" group the magnitude of RMSE was large by comparison, indicating that there was less accurate recovery of the ICCs throughout the range restriction conditions. For the "B" group, which was assumed to have taken a test that contained either 10 biased items or 40 biased items, the values of RMSEs under Z and N conditions were below 0.06 across the sample sizes except for the Z

Condition with N=600. These RMSE values indicate quite accurate recovery of ICCs in the study, although it is apparent that range restriction affected parameterization.

CONCLUSIONS

The major purpose of the present study was to investigate how range restriction in the ability distribution of the sample influences the detection accuracy of the several item bias detection techniques. It is concluded that for a test with 10 biased items, range restriction had a clearly systematic influence on the accuracy of three of the bias detection techniques. These include total area index, base high area index and chi-square index. The correlations between the generated bias and the detected bias by these techniques exhibited a steady decrease with an increase in the degree of range restriction in the sample for the larger sample sizes. When these techniques were applied to the test that had 40 biased items, the effects of range restriction were less systematic.

The Pearson correlations between the true bias and detected bias with Base Low Area index were close to zero across all of the restriction-sample size Combinations. This was true for both the test including 10 biased items and 40 biased items, except for a decreasing trend in the correlations with an increase in the degree of restriction when both groups had the same number of examinees and were simulated to take the test with 40 biased items.

The Signed Area Index seemed to be robust to range restriction in the ability of sample, and was generally accurate in determining biased items even under the severest range restriction conditions, and this did not vary as a function of the number of biased items in a test.

When sample sizes for two groups were equal, the RMSD index was sensitive to ability range restriction in two groups. In contrast there was little indication that the degree of range restriction in the group samples had a role in the degree of accuracy of the index when the groups compared had different numbers of examinees.

When the results based on 50 items whose difficulty parameter values ranged between -3 and +3 were compared with the results based on only those 27 items whose item difficulty parameters ranged between -2 and +2, the effect of range restriction tended to be much clearer with the 27 item test. This was expected because previous researchers have indicated that when a test does not have very difficult or very easy items, estimation of IRFs tended to be very accurate. Accurate estimation of the item response functions in turn might lead to more precise results of the effect of range restriction on the bias detection indices. When N=1000 for both groups, the Pearson correlations between the simulated and detected bias tended to be lower when there was no range restriction as compared to the Z restriction condition where examinees were randomly sampled above -1.0 from the normal distribution. To examine the possibility of random sampling error, 1,000 ability values were randomly sampled for the "A" group (i.e., the no restriction conditions were replicated for N=1000 in both groups) and the bias detection indices including Total, Base High, Signed Area, Base Low and Chi square were recalculated. The new values for Total index, Base High, Base Low, Signed Area, and Chi Square were 0.34, 0.55, 0.09, 0.59, 0.70 respectively. The recalculated correlations were larger than the original ones, and a little larger than those for the Z Conditions presented in Tables 1-4, and Table 6. It might be concluded that recovery of ICCs was poorer below -1.0 in

the initial sample, and that having no restriction in the range of the sample is not always best condition to detect item bias, and that having some range restriction may still enable one to obtain reasonably accurate detection of item bias by most of the bias detection indices. This may be somewhat consistent with the fact that persons in the extremes of the ability continuum have abilities that are likely to be more poorly estimated. Removing these cases can sometimes make the overall calibration more accurate.

Root Mean Square Errors were computed to cross validate the results obtained for the bias detection techniques. Recovery of the ICCs became worse as range restriction increased regardless of the amount of item bias. This might have led to the expectation that, since error in estimating ICCs shows a clearcut and systematic increment with an increase in the degree of restriction, effect of restriction on the bias detection indices would have been clear. However, this was not always the case and may be attributed to the observation that the accuracy of recovery of ICCs was not the same on the different regions of the ability continuum.

One limitation of the present study is that restrictions were created by directly truncating the normal distribution and using the upper part of the distribution. It should be noted that when the ability values and the number right scores that were based on simulated responses were correlated for the various samples used in the study, the correlations were on average approximately 0.85. In certain applied settings it may well be that the ability (θ) distribution for test-takers is skewed rather than truncated, since most organizations select people on the basis of raw scores. Such restriction is likely to be incidental, and not direct as in the present simultaion.

REFERENCES

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

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Berk, R. A. (1982) Introduction. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Burrill, L. E. (1982) Comparative studies of item bias methods. In R. A. Berk (Ed.) Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Drasgow, F. (1982) Biased test items and differential validity. Psychological Bulletin, 92, 526-531.

Drasgow, F. (1984) Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. Psychological Bulletin, 95, 134-135.

Ghiselli, E. E., Campbell J. P., & Zedeck, S. (1981) Measurement theory for the behavioral sciences. San Francisco: W.H. Freeman.

Green, D. R., & Draper, J. F. (1972) Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September. (ERIC Document Reproduction Service No. ED 070 794)

Guion, R. M., & Ironson, G. (1983) Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31, 54-87.

Hambleton, R. K. & Swaminathan H. (1985) Item response theory: principles and applications. Hingham, MA: Kluwer-Nijhoff.

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.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983) Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Ironson, G .H. (1982) Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods of detecting test bias. Baltimore: Johns Hopkins University Press.

Linn, R. L., & Harnish, 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 comprehesion. Applied Psychological Measurement, 5, 159-163.

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

McCauley, C. D. & Mendoza, J. L. (1985) A simulation study of item bias using a two parameter item-response model. Applied Psychological Measurement, 9, 389-400.

Nungster, R. J. (1977) An empirical examination of three models of item bias. Dissertation Abstracts International, 38, 2726A.

Petersen, N. S. (1977) Bias in the selection rule: Bias in the test. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, Netherlands.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3-29.

Ree, M. J. (1979) Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Rudner, L. M. (1977) An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University of America, 1977.

Shepard, L. A. Definition of bias (1982) In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

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

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

Swaminathan, H., & Gifford, J. A. (1980) Estimation of parameters in the three parameter latent trait model. Laboratory of Psychometric and Evaluation Research (Report No. 90) Amherst, Mass: University of Massachusetts, School of Education.

Wingersky, M. S., Barton, M. A., & LORD. F. M. (1982) Logist User's guide. Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977) Solving Measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116

Wright, B. D., Mead, R. J., & Draba R. (1976) Detecting and correcting test item bias with a logistic response model (Research Memorandom no. 22) Chicago: Statical Laboratory, Department of Education, University of Chicago. Wright, B. D., & Stone, M. H. (1979) Best test design. Chicago: MESA press, 1979.

Footnotes

[1] This exception was necessitated by the fact that there is no sampling error within the entire subpopulation, so a true χ^2 could not be calculated. As the total area index takes into account both slope and difficulty differences, it was chosen as representing similar information as that captured in the χ^2 index. It should be noted that previous research has generally found the highest correlations of the χ^2 index with the Total Area index (Shepard, et al., 1981; McCauley & Mendoza, 1985).

Table 1

Correlations Between True and Detected Bias: Total Area Index

Condition	10 Biased Items		40 Biased Items	
	50 items	27 items	50 items	27 items
A (n=1000)				
W B (n=1000)	0.05	0.14	0.35	0.15
B (n=600)	0.16	0.32	0.32	0.27
B (n=300)	0.45	0.74	0.49	0.58
X B (n=1000)	0.26	0.30	0.37	0.28
B (n=600)	0.16	0.35	0.27	0.04
B (n=300)	0.06	0.09	0.33	0.27
Y B (n=1000)	0.36	0.68	0.43	0.47
B (n=600)	0.27	0.49	0.28	0.36
B (n=300)	0.05	0.15	0.43	0.44
Z B (n=1000)	0.34	0.72	0.49	0.76
B (n=600)	0.23	0.75	-0.07	0.26
B (n=300)	0.32	0.80	0.45	0.63
N B (n=1000)	0.19	0.58	0.07	0.16
B (n=600)	0.30	0.85	0.20	0.16
B (n=300)	0.25	0.53	0.10	0.21
A (n=600)				
W B (n=600)	0.12	0.25	0.10	0.25
B (n=300)	0.39	0.54	0.32	0.43
X B (n=600)	0.19	0.49	0.30	0.23
B (n=300)	0.04	0.14	0.34	0.24
Y B (n=600)	0.49	0.71	0.61	0.64
B (n=300)	0.26	0.45	0.40	0.36
Z B (n=600)	0.21	0.56	0.34	0.46
B (n=300)	0.39	0.69	0.39	0.38
N B (n=600)	0.32	0.84	0.16	0.24
B (n=300)	0.28	0.62	0.05	0.27
A & B (n=300)				
W	-0.03	0.14	0.04	0.08
X	0.09	0.12	0.38	0.30
Y	0.04	0.09	0.25	-0.05
Z	0.01	0.36	0.25	0.22
N	0.54	0.82	0.13	0.24

Note. "W" was most severe range restriction condition. "N" was no range restriction condition. "A" indicates majority group, "B" indicates minority group.

Table 2

Correlations Between True and Detected Bias: Base High Area Index

Condition	10 Biased Items		40 Biased Items	
	50 items	27 items	50 items	27 items
A (n=1000)				
W B (n=1000)	0.16	0 51	0.20	0.22
B (n=600)	0.16	0.49	0.19	0.22
B (n=300)	0.18	0.47	0.07	0.15
X B (n=1000)	0.22	0.43	0.22	0.27
B (n=600)	0.12	0.52	0.13	0.19
B (n=300)	0.06	0.38	0.14	0.16
Y B (n=1000)	0.43	0.81	0.34	0.47
B (n=600)	0.36	0.68	0.21	0.27
B (n=300)	0.20	0.41	0.38	0.46
Z B (n=1000)	0.54	0.89	0.67	0.92
B (n=600)	0.47	0.81	0.17	0.30
B (n=300)	0.54	0.93	0.66	0.81
N B (n=1000)	0.46	0.84	0.38	0.59
B (n=600)	0.52	0.95	0.46	0.58
B (n=300)	0.48	0.77	0.38	0.66
A (n=600)				
W B (n=600)	0.18	0.44	0.16	0.25
B (n=300)	0.16	0.37	0.08	0.15
X B (n=600)	0.14	0.53	0.14	0.22
B (n=300)	0.07	0.38	0.14	0.14
Y B (n=600)	0.36	0.58	0.47	0.58
B (n=300)	0.22	0.40	0.43	0.50
Z B (n=600)	0.39	0.69	0.51	0.67
B (n=300)	0.57	0.90	0.55	0.63
N B (n=600)	0.41	0.94	0.39	0.58
B (n=300)	0.38	0.81	0.33	0.66
A & B (n=300)				
W	0.06	0.19	0.09	0.21
X	0.02	0.30	0.12	0.10
Y	0.06	0.37	0.15	0.13
Z	0.32	0 82	0.37	0.45
N	0.62	0.86	0.39	0.67

Note. "W" was most severe range restriction condition. "N" was no range restriction condition. "A" indicates majority group, "B" indicates minority group.

Table 3

Correlations Between True and Detected Bias: Base Low Area Index

Condition	10 Biased Items		40 Biased Items	
	50 items	27 items	50 items	27 items
A (n=1000)				
W B (n=1000)	-0.15	-0.22	0.28	-0.28
B (n=600)	-0.14	0.06	0.26	-0.20
B (n=300)	-0.02	0.18	0.01	-0.22
X B (n=1000)	-0.04	-0.27	0.36	-0.05
B (n=600)	0.01	-0.17	0.46	0.06
B (n=300)	-0.02	-0.22	0.28	0.16
Y B (n=1000)	0.09	-0.01	0.55	0.35
B (n=600)	0.03	-0.16	0.07	-0.11
B (n=300)	-0.07	0.07	0.39	0.38
Z B (n=1000)	0.00	-0.19	0.44	0.47
B (n=600)	-0.00	-0.08	0.25	0.04
B (n=300)	0.22	0.13	0.66	0.49
N B (n=1000)	0.06	-0.12	0.50	0.19
B (n=600)	0.03	0.08	0.53	0.16
B (n=300)	0.06	0.09	0.39	0.23
A (n=600)				
W B (n=600)	-0.22	-0.19	0.17	0.06
B (n=300)	-0.09	-0.13	-0.11	-0.07
X B (n=600)	0.08	0.47	0.56	0.16
B (n=300)	0.02	0.21	0.36	0.18
Y B (n=600)	-0.11	-0.17	0.11	-0.08
B (n=300)	-0.19	-0.15	0.32	0.14
Z B (n=600)	0.02	-0.10	0.38	0.11
B (n=300)	0.26	0.14	0.38	0.20
N B (n=600)	0.46	0.16	0.50	0.25
B (n=300)	0.03	0.15	0.39	0.29
A & B (n=300)				
W	-0.18	-0.26	-0.05	-0.09
X	-0.15	-0.15	0.27	0.12
Y	0.17	0.08	0.33	-0.16
Z	0.08	0.14	0.47	0.17
N	-0.06	-0.01	0.29	0.01

Note. "W" was most severe range restriction condition. "N" was no range restriction condition. "A" indicates majority group, "B" indicates minority group.

Table 4

Correlations Between True and Detected Bias: Signed Area Index

Condition	10 Biased Items		40 Biased Items	
	50 items	27 items	50 items	27 items
A (n=1000)				
W B (n=1000)	0.42	0.61	0.65	0.71
B (n=600)	0.41	0.64	0.68	0.71
B (n=300)	0.51	0.70	0.47	0.63
X B (n=1000)	0.59	0.71	0.62	0.64
B (n=600)	0.49	0.76	0.56	0.66
B (n=300)	0.37	0.61	0.50	0.53
Y B (n=1000)	0.55	0.75	0.43	0.70
B (n=600)	0.49	0.67	0.45	0.49
B (n=300)	0.35	0.45	0.70	0.74
Z B (n=1000)	0.55	0.84	0.74	0.95
B (n=600)	0.46	0.78	0.35	0.31
B (n=300)	0.57	0.85	0.70	0.83
N B (n=1000)	0.54	0.77	0.63	0.72
B (n=600)	0.56	0.86	0.64	0.72
B (n=300)	0.55	0.75	0.54	0.72
A (n=600)				
W B (n=600)	0.51	0.64	0.59	0.64
B (n=300)	0.53	0.67	0.39	0.55
X B (n=600)	0.45	0.76	0.54	0.59
B (n=300)	0.41	0.64	0.45	0.44
Y B (n=600)	0.60	0.80	0.63	0.68
B (n=300)	0.45	0.61	0.69	0.69
Z B (n=600)	0.50	0.65	0.69	0.70
B (n=300)	0.60	0.76	0.69	0.67
N B (n=600)	0.56	0.88	0.59	0.70
B (n=300)	0.54	0.77	0.53	0.69
A & B (n=300)				
W	0.11	0.19	0.10	0.26
X	0.30	0.56	0.42	0.60
Y	0.28	0.40	0.46	0.48
Z	0.38	0.69	0.51	0.64
N	0.68	0.78	0.54	0.65

Note. "W" was most severe range restriction condition. "N" was no range restriction condition. "A" indicates majority group, "B" indicates minority group.

Table 5

Correlations Between True and Detected Bias: RMSD Index

Condition	10 Biased Items		40 Biased Items	
	50 items	27 items	50 items	27 items
A (n=1000)				
W B (n=1000)	0.04	0.15	0.49	0.33
B (n=600)	0.18	0.42	0.45	0.43
B (n=300)	0.34	0.74	0.58	0.70
X B (n=1000)	0.30	0.29	0.42	0.35
B (n=600)	0.13	0.35	0.34	0.16
B (n=300)	0.11	0.08	0.44	0.37
Y B (n=1000)	0.39	0.70	0.52	0.65
B (n=600)	0.25	0.47	0.23	0.36
B (n=300)	0.03	0.14	0.53	0.58
Z B (n=1000)	0.31	0.71	0.51	0.80
B (n=600)	0.15	0.77	-0.12	0.28
B (n=300)	0.30	0.79	0.49	0.73
N B (n=1000)	0.16	0.47	0.03	0.10
B (n=600)	0.32	0.70	0.16	0.16
B (n=300)	0.29	0.54	0.01	0.19
A (n=600)				
W B (n=600)	0.14	0.30	0.20	0.31
B (n=300)	0.33	0.51	0.41	0.50
X B (n=600)	0.10	0.50	0.29	0.29
B (n=300)	0.00	0.03	0.39	0.29
Y B (n=600)	0.37	0.64	0.56	0.60
B (n=300)	0.19	0.40	0.47	0.46
Z B (n=600)	0.13	0.46	0.29	0.54
B (n=300)	0.40	0.63	0.44	0.54
N B (n=600)	0.30	0.80	0.12	0.25
B (n=300)	0.25	0.68	-0.05	0.24
A & B (n=300)				
W	-0.05	0.10	0.05	0.14
X	0.06	-0.02	0.48	0.46
Y	0.10	0.18	0.37	0.20
Z	0.04	0.30	0.23	0.24
N	0.55	0.83	0.04	0.27

Note. "W" was most severe range restriction condition. "N" was no range restriction condition. "A" indicates majority group, "B" indicates minority group.

Table 6

Correlations Between True and Detected Bias: Lord's Chi-Square Index

Condition	10 Biased Items		40 Biased Items	
	50 items	27 items	50 items	27 items
A (n=1000)				
W B (n=1000)	0.11	0.10	0.21	-0.07
B (n=600)	0.06	0.10	0.27	0.29
B (n=300)	0.05	0.87	0.23	0.36
X B (n=1000)	0.54	0.53	0.38	0.13
B (n=600)	0.45	0.54	0.21	-0.06
B (n=300)	0.17	0.23	0.24	0.14
Y B (n=1000)	0.56	0.64	0.46	0.37
B (n=600)	0.55	0.61	-0.01	0.03
B (n=300)	0.35	0.52	0.32	0.24
Z B (n=1000)	0.54	0.73	0.46	0.53
B (n=600)	0.45	0.71	0.19	0.25
B (n=300)	0.45	0.72	0.40	0.44
N B (n=1000)	0.61	0.79	0.28	0.30
B (n=600)	0.71	0.92	0.19	0.18
B (n=300)	0.48	0.65	0.40	0.19
A (n=600)				
W B (n=600)	-0.06	-0.22	0.27	0.35
B (n=300)	-0.05	-0.12	0.11	0.11
X B (n=600)	0.53	0.67	0.25	0.21
B (n=300)	0.22	0.23	0.33	0.27
Y B (n=600)	0.59	0.77	0.16	0.10
B (n=300)	0.53	0.67	0.35	0.32
Z B (n=600)	0.53	0.63	0.56	0.55
B (n=300)	0.50	0.59	0.39	0.36
N B (n=600)	0.80	0.94	0.32	0.26
B (n=300)	0.59	0.72	0.26	0.29
A & B (n=300)				
W	-0.07	-0.18	0.15	0.21
X	0.10	0.08	0.38	0.22
Y	-0.22	-0.34	0.30	0.12
Z	0.34	0.43	0.40	0.34
N	0.52	0.57	0.22	0.20

Note. "W" was most severe range restriction condition. "N" was no range restriction condition. "A" indicates majority group, "B" indicates minority group.

Table 7

Average Root Mean Square Errors
Across 50 items

Sample Size	R	Group A	Group B(10)	Group B(40)
1000	W	0.2909	0.1328	0.1263
	X	0.3138	0.1207	0.1428
	Y	0.2533	0.0932	0.1046
	Z	0.1774	0.0591	0.0554
	N	0.1200	0.0550	0.0419
600	W	0.3286	0.1318	0.1509
	X	0.3112	0.0777	0.1246
	Y	0.2638	0.0638	0.1176
	Z	0.1712	0.1076	0.0530
	N	0.1109	0.0387	0.0390
300	W	0.3057	0.2329	0.1998
	X	0.3113	0.0921	0.1218
	Y	0.3230	0.0970	0.0727
	Z	0.1950	0.0575	0.0560
	N	0.1387	0.0491	0.0570

Note. Column R denotes the restriction condition with "W" being the most severe and "N" being no range restriction. Group B(10) denotes a test containing 10 biased items. Group B(40) denotes a test containing 40 biased items.