Improving IRT Item Bias Detection With Iterative Linking and Ability Scale Purification

Dong-Gun Park and Gary J. Lautenschlager
University of Georgia

The effectiveness of several iterative methods of item response theory (IRT) item bias detection was examined in a simulation study. The situations employed were based on biased items created using a two-dimensional IRT model. Previous research demonstrated that the non-iterative application of some IRT parameter linking procedures produced unsatisfactory results in a simulation study involving unidirectional item bias. A modified form of Drasgow's iterative item parameter linking method and an adaptation of Lord's test purification procedure were examined in conditions that simulated unidirectional and mixed-directional forms of item bias. The results illustrate that iterative linking holds promise for differentiating biased from unbiased items under several item bias conditions. In addition, a combination of methods, involving cycles of iterative linking followed by ability scale purification, was found to be even more effective than iterative linking alone. This combination of procedures totally eliminated false positive misidentifications for the most pervasive item bias condition, and false negative misidentifications were also reduced. Combining iterative linking with ability scale purification appears to be a viable method for analyzing multidimensional IRT data with unidimensional IRT item-bias methods. *Index terms: ability scale purification, item bias, item response theory, iterative linking, iterative methods, metric linking, multidimensional IRT model.*

Research related to the detection of item bias has proliferated in the psychometric and applied psychological literature (Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Williams, 1985). Among item bias methods, the theoretically-preferred method is based on item response theory (IRT). It is preferred because of its sample invariant properties, which make it less likely that true subpopulation differences will be mistaken for bias. Among others, Berk (1982, p. 3) and Drasgow (1982, 1984) have provided a sound justification for more direct concern with item bias detection, rather than with overall predictive bias as determined from total test scores.

Two notions are central to the concept of item bias. The first notion is that examinee performance on an item may be influenced by sources of variation other than differences on the dimension of interest. The second is that these extraneous sources of variation affect performance in a way that differs systematically for some subpopulations, which gives an unfair advantage to one subpopulation over another. Based on these two notions, a definition of an unbiased item can be formulated: An item is unbiased with respect to two subpopulations if the item is influenced by the same sources of variance in both subpopulations; in addition, among examinees who are at the same level on the dimension purportedly measured by the test, the distributions of irrelevant sources of variation are the same for both subpopulations (Crocker & Algina, 1986).

The IRT method for detecting item bias is based on the comparison of item response functions (IRFs) estimated separately for two groups. If an item is unbiased, then the IRFs for different groups should be the same. If IRFs for two groups differ by more than sampling error, then the item is suspected of being biased.

Metric Linking and IRT Item Bias Analysis

Lautenschlager and Park (1988) recently demonstrated that the linking of item parameter and person metrics in IRT item bias analysis is

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 2, June 1990, pp. 163–173

0146-6216/90/020163-11$1.80

not without its own potential pitfalls. This is because all the IRT item bias detection methods compare the IRFs obtained for the separate groups. The linking of person ability metrics occurs through a standardization process involving item parameter estimates, and a number of methods have been developed that could ostensibly serve this purpose (Divgi, 1985; Linn, Levine, Hastings, & Wardrop, 1981; Lord, 1980; Stocking & Lord, 1983; Warm, 1978). However, there is a paradox involved in IRT item bias analysis: The common items that should be involved in linking ability (θ) metrics are the truly unbiased items, the very items that can be identified through the item bias analysis itself.

The more sophisticated metric linking methods, such as those offered by Stocking and Lord (1983) and Divgi (1985), might be even less appropriate for item bias analyses because they use more of the information available from biased items, as well as unbiased items, in order to achieve the goal of symmetric treatment of the two sets of item parameter estimates. For example, Stocking and Lord (1983) found their characteristic curve method better achieved this goal by providing a better fit. This fit was defined as bisection of the plots of a given set of item parameter estimates from two calibration samples in a different type of linking situation in which the samples may safely be presumed to come from the same parent population (Stocking & Lord, p. 205–207). However, this might not necessarily be appropriate for item bias research in which the plotted points likely represent uncommon (i.e., biased), as well as common (i.e., unbiased), items. In theory, if a test has biased items—in particular, many biased items that are biased against one group—the linking line should not be expected to bisect the item parameter estimate point cloud from two independent calibration samples.

Improving IRT Item Bias Detection Methods

Lord (1980) described a procedure (that is attributed to Marco, 1977) called "purification," as having some potential for alleviating this linking dilemma.

Lord (1980, p. 220) described the procedure as follows:

> 1. Analyze the total test, as described in the preceding sections. [These "preceding sections" refer to estimating item parameters for all groups combined and standardizing on the b_i parameters; fixing c_i parameters to those obtained in that analysis; and then re-estimating a_i and b_i within each group, again standardizing on b_i; finally, IRFs are compared for evidence of bias. See Lord, 1980, pp. 213–220.]
> 2. Remove all items that have significantly different response functions in the groups under study. The remaining items may now be considered to be a unidimensional pool, even when the groups are combined.
> 3. Combine all groups and estimate θ for each individual. These θs should all be comparable.
> 4. For each group separately, while holding θ fixed for all individuals at the values obtained in step 3, estimate the a_i and the b_i for each item. Do this for all items, including those previously removed.
> 5. Compare estimated item response functions or parameters by the methods of section 14.4 [computation of a χ^2 test of differences in IRFs]. (p. 220)

The rationale is that if many items are found to be seriously biased, then it appears that the items are not strictly unidimensional. The θ estimates obtained for one group are not strictly comparable to the θ estimates obtained for another. This casts some doubt on the results obtained when all items are analyzed together.

This purification procedure does not directly address the linking issue, but rather considers the dimensionality of the data. Because multidimensionality and the linking issue are in some sense related to each other, however, this procedure could suggest a possible way of overcoming the linking dilemma. By removing potentially biased

items, the procedure attempts to obtain a unidimensional latent space; thus it could aid in resolving the scale-linking dilemma noted earlier. Although it has some potential, there is little indication that this procedure has been employed in published articles, except in Lord's (1977) item bias study.

More recently, Drasgow (1987) used a procedure that has potential utility for overcoming the linking problem. The procedure involved the application of the Stocking and Lord (1983) metric linking method to initially place item parameter estimates on approximately equal scales. Next, item bias statistics were computed. Subsequently, item parameter estimates were relinked, using only those items found to be unbiased in the previous step; again, item bias statistics were computed for all the items. This process continued until the same set of items was found to be biased on two successive iterations. The rationale for using the iterative procedure was that if biased items were not discarded, linking methods might compensate for truly biased items by causing unbiased items to appear biased.

This procedure could have more potential than Lord's purification, but it is not without problems. For example, if there were many biased items in a test, or the magnitude of item bias were large, initial estimates of person and item parameters might not be accurate. In this case, the iterative linking used by Drasgow would not be effective because the procedure would continue to use the potentially inaccurate estimates obtained in the initial estimation. Thus after eliminating the biased items identified on initial linking, it might be necessary to re-estimate person and item parameters, using only the items detected as being "unbiased."

Another alternative method for identifying biased items was adapted by Park (1988) from the test purification procedure described by Lord (1980), and is referred to as the modified-Lord test purification (M-LTP) method. This procedure functions as follows:

1. Combine all groups, and estimate θ for each individual.
2. For each group separately, while holding θ estimates fixed for all examinees at the values obtained in step 1, estimate the item parameters and compute the item bias statistic for each item.
3. Remove all items that have significantly different response functions in the groups.
4. Using the remaining items from step 3, estimate θ for each individual.
5. Estimate the item parameters for all the items in the test for each group separately, while holding θ estimates fixed for all individuals at the values obtained in step 4.
6. Compute item bias statistics for each item in the test.
7. Repeat this process (steps 3 through 6) until the same set of items is found to be biased on two successive iterations.

A potential virtue of this procedure is that no metric linking method is explicitly involved in the item bias detection procedure. This should eliminate concern about the statistical artifacts that may result from use of a metric linking method, thus possibly overcoming the linking dilemma to some degree. The present study was designed to compare the two iterative item bias detection procedures described above, and to examine their potential to resolve the linking dilemma (Lautenschlager & Park, 1988).

Method

Simulation of Item Responses

The simulation of item response data for examinees was based on variations of the three-parameter logistic model. This model was chosen because it has been found to produce a realistic reflection of data from standardized achievement tests (Ansley & Forsyth, 1985).

Previous monte carlo simulation studies of item bias that have employed unidimensional IRT item bias detection methods (e.g., Drasgow, 1987; Rudner et al., 1980; Shepard et al., 1985) have increased or decreased the values of item parameters for one of the groups being compared, typically by shifting item difficulty

parameters for a subset of items. McCauley and Mendoza (1985), however, used a factor analytic model for developing biased item responses.

An alternative method of creating biased items, which allows for additional sources of variance (θ dimensions) to systematically influence item performance, is more consonant with the IRT definition of biased items. The model adopted here for simulating biased item responses was that used by Park (1988; Lautenschlager & Park, 1988), in which biased items were items that were unidimensional for one group, while some of these same items were multidimensional in the other group (Hambleton & Swaminathan, 1985). This model involved the use of an incidental θ dimension to influence performance on biased items. Such an operational definition of item bias involved a focal dimension (θ_1) and a second incidental dimension (θ_2) that could influence performance on some items.

A two-dimensional version of the multidimensional IRT model proposed by Sympson (1978) was used to generate biased item responses. The Sympson non-compensatory model is

$$P_{ij}(\theta_{ih}) = c_j + \frac{1 - c_j}{\Pi \{1 + \exp[-1.7a_{jh}(\theta_{ih} - b_{jh})]\}} \quad (1)$$

where:

θ_{ih} is the ability parameter for person i on dimension h,

a_{jh} is the discrimination parameter for item j on dimension h,

b_{jh} is the difficulty parameter for item j on dimension h, and

c_j is the pseudo-guessing parameter for item j.

Ansley and Forsyth (1985) justified the use of Sympson's non-compensatory model in preference to other models by indicating that the non-compensatory view of dimensionality is more reasonable for most well-constructed achievement tests. In addition, the model produced data that had properties similar to actual achievement test data.

The unidimensional three-parameter logistic model was used here to generate data for all unbiased items. The θ dimensions employed are referred to as θ_1 (the focal dimension common to all items and groups) and an incidental dimension, which is referred to as θ_2; the latter dimension produced item bias as described below.

The rationale and guidelines for selecting item parameters and generating item response data to reflect actual test data are described in detail elsewhere (Lautenschlager & Park, 1988). Essentially, item difficulty parameters associated with the θ_1 dimension were sampled from a uniform distribution in the interval from −2.0 to +2.0, and item discrimination parameters were sampled from a uniform distribution of .6 to 2.0 (Swaminathan & Gifford, 1980). For the incidental dimension θ_2, item difficulty parameters were scaled to have a mean of −1.0 and a standard deviation of .70. Item discrimination values for θ_2 were centered at .50, with a standard deviation of about .10. The c_j parameters were held constant at .20.

Dataset Characteristics

Datasets were generated to simulate item responses to multiple-choice items with four response options. Data were simulated for two groups, Group A and Group B. All simulated datasets had 1,000 examinees in each group. The number of biased items in a given simulated test was varied, resulting in either 18, 28, or 36 biased items out of a total of 54 items in the test. The inclusion of pervasive amounts of biased items permitted an examination of the robustness of IRT parameter estimates and item bias detection procedures.

Six pairs of datasets were created based on combinations of the number of biased items, $\hat{\theta}$ distributions on the first and second dimensions, correlation of the latent dimensions, and direction of bias. Four of these pairs were involved in the simulation of unidirectional bias conditions, in which bias was against Group B, and the other two pairs simulated mixed directional bias conditions. These datasets are described in detail be-

Table 1
Number of Unbiased and Biased Items, Mean (M) and Standard Deviation (SD) of the Incidental Trait (θ_2), and the Population Correlation Between the Focal and Incidental Trait [$r(\theta_1\theta_2)$] in Unidirectional and Mixed Directional Item Bias Conditions (N = 1,000)

Bias Condition	Group	Number of Unbiased Items	Number of Biased Items	Normal θ_2 Distribution M	SD	$r(\theta_1\theta_2)$
Unidirectional						
1	A	54	0	-	-	-
	B	36	18	0	1.0	.90
2	A	54	0	-	-	-
	B	36	18	-.5	1.0	.90
3	A	54	0	-	-	-
	B	26	28	-.5	1.0	.60
4	A	54	0	-	-	-
	B	18	36	-.5	1.0	.60
Mixed Directional						
5	A	36	18	-.5	1.0	.60
	B	36	18	-.5	1.0	.60
6	A	36	18	0	1.0	.90
	B	36	18	-.5	1.0	.90

low. Table 1 summarizes the characteristics of the datasets generated to represent both the four unidirectional bias conditions, and the two mixed item bias conditions.

Unidirectional Bias Conditions. For the unidirectional bias conditions, it was assumed that only θ_1 influenced performance on all unbiased items. Thus the three-parameter logistic IRT model was used to generate item response data for all items for Group A examinees, and for all unbiased items for Group B examinees. The generation of item responses for the biased items involved the use of the two-dimensional version of Sympson's model. Two types of normal θ distributions on θ_2 were generated for the B group examinees using a mean of either -.5 or 0, and a standard deviation of 1.0. The correlation between the focal (θ_1) and incidental (θ_2) dimensions was either .60 or .90 in the population.

Mixed Directional Bias Conditions. For the mixed directional bias conditions, it was assumed that a separate, unrelated (across groups) incidental θ influenced performance on biased items within each of the two groups. Consequently, two additional sets of item parameters and two distinct incidental θ dimensions for biased items were generated in a given mixed bias condition. Items biased against one group were unbiased in the other group. The generation of biased items proceeded as it had in the unidirectional bias conditions, except that biased items were created for both groups. Two types of normal θ_2 distributions were generated again using a mean of either 0 or -.5, and all such distributions had a standard deviation of 1.0 in common. The correlation of θ_1 with θ_2 in the population was set at either .60 or .90, and was the same for both groups within a given bias condition.

Analysis

The LOGIST computer program (Wingersky, Barton, & Lord, 1982) was used to estimate item and θ parameters, and was used in successive iterations for estimation with fixed estimated θ values for the M-LTP procedure described earlier. Lord's (1980) χ^2 item bias statistic was used to indicate the potential for item bias, and a significance level of .005 was used for indicating "detected" item bias. Due to the costs involved with using the M-LTP item bias detection proce-

dures, only bias conditions 1, 3, 4, and 6 were examined for this method.

Divgi's (1985) minimum χ^2 method was substituted in place of the more complex Stocking and Lord (1983) method in the linking of estimates of IRT parameters, producing a modification of Drasgow's (1987) iterative linking procedure (M-DIL). This method was chosen because it was simpler and less expensive than other methods, and it had desirable features found in more complex methods, such as that of Stocking and Lord (1983). Thus the M-DIL procedure employed a different metric linking method than that used by Drasgow (1987). Relinking of parameters for the M-DIL procedure was accomplished by applying Divgi's method to subsets of items that had been flagged as unbiased in the immediately preceding iteration in order to determine linking constants.

Results

Table 2 presents the number of false positives (FPs) and false negatives (FNs) on each iteration for all test bias conditions investigated.

Unidirectional Test Bias Conditions

For bias conditions 1 and 2, each with eighteen truly biased items, the difference in the mean of the θ_2 distributions between these two conditions did not greatly affect the outcome when M-DIL was used. Both conditions converged in just three iterations, with the same set of items being identified as biased in the second and third iterations. There were no FPs identified in either condition, because all truly unbiased items were identified as being unbiased on the final iteration. Indeed, FPs were rare regardless of iteration. Eight FNs were identified in condition 1 for both methods and seven FNs occurred in condition 2. The use of the M-LTP method for bias condition 1 produced the same outcome as the M-DIL method, except that more iterations were needed to reach convergence with the former method. False negatives tended to occur among the more weakly biased items, as might be expected.

In bias condition 3 [$\theta_2 \sim N(-.5, 1.0)$ and $r(\theta_1,\theta_2) = .6$], 28 items were created to be biased. Here the number of iterations required by the M-DIL method to reach convergence increased to six. Only 4 out of 11 FNs were later correctly detected as biased items on the final iteration. All 9 FPs on the first iteration were correctly identified as unbiased by the final iteration. For the M-LTP method, the number of FP identifications in the initial stage was 19 out of 26 possible. One-half of the truly biased items resulted in FN misclassifications. This was substantially more degenerate than the results for the M-DIL method, and suggested that the M-LTP method likely would not lead to better results in further iterations; therefore, no further iterations were done with the M-LTP method (partly due to cost considerations, as well).

In bias condition 4, with the most pervasive unidirectional item bias, both the M-DIL and M-LTP methods led to less than desirable solutions. Although convergence was reached in four iterations for the M-DIL method, one-half of the truly unbiased items were incorrectly identified as biased, and nearly half of the truly biased items were identified as unbiased. As had happened for bias condition 3, the use of the M-LTP method produced poor results on the first iteration; thus no further iterations of the M-LTP method were attempted for this condition.

It could be argued that the test simulated in bias condition 4, which was intended to contain 36 biased items, was actually more accurately a test with only 20 biased items. Because the tests with 28 biased items (in bias condition 3) contained seven false negatives that also could have been weakly biased, that test could be taken for a test with 21 biased items. It is tempting then to conclude that the two tests represented by conditions 3 and 4 were similar in terms of degree of bias. However, if the two tests were biased to approximately the same degree, why should using the M-DIL method alone have resulted in no FPs at convergence in condition 3, while detecting nine FPs in condition 4 after convergence? This result was probably not solely due to sam-

Table 2
False Positives (FP) and False Negatives (FN) Observed on Each Iteration for the Unidirectional and Mixed Test Bias Conditions, for Modified Drasgow Iterative Linking Method (MD) and the Modified Lord Test Purification Method (ML)

		Unidirectional							Mixed		
		1		2	3		4		5	6	
Iteration		MD	ML	MD	MD	ML	MD	ML	MD	MD	ML
1											
	FP	1	1	0	9	19	11	18	0	0	1
	FN	10	10	8	11	14	17	12	2	2	3
2											
	FP	0	0	0	2		9		0	0	0
	FN	8	9	7	11		17		3	2	4
3											
	FP	0	0	0	0		9		0		0
	FN	8	8	7	10		16		3		2
4											
	FP		0		0		9				0
	FN		8		8		16				3
5											
	FP				0						
	FN				7						
6											
	FP				0						
	FN				7						

pling error. One possible explanation is that although each of the 16 FNs (prior to the purifications) were weakly biased, the cumulative effect for those 16 FNs was large enough to distort the θ scale and therefore produce many false positives.

At this point a post hoc procedure was implemented to determine if a variation involving features of both iterative parameter linking and test purification procedures could improve the correct determination of biased and unbiased items for bias condition 4. Rather than use subsequent purification of the θ scales immediately after each iteration, such as in the M-LTP method, scale purification was done after initial convergence had been achieved using the M-DIL procedure. This amounted to using both iterative parameter linking and θ scale purification (ILAP).

Application of the ILAP procedure involved starting with the results at convergence of the M-DIL method, as described in the last iteration of condition 4 from Table 2. From these results, θs were re-estimated separately within each group, using only those items which had been flagged as "unbiased" from that last M-DIL iteration. Holding these newly estimated θs as fixed, item parameters were re-estimated for all items in the test, and again this was done separately within each group. Using Divgi's method, linking constants were estimated, based on only those items that were flagged as "unbiased" preceding the purification step, and then new item bias statistics were calculated. This analysis produced 2 FPs and 15 FNs. The M-DIL method was applied once more, and produced no changes in item classifications on convergence at the second iteration.

Because the purification had resulted in changes in item classifications, a second pass of this entire procedure was conducted using data from the last M-DIL step noted above as a starting point. The second θ scale purification step resulted in 1 FP and 14 FNs, once item bias statistics were calculated. After convergence on the third iteration, application of the M-DIL method

produced 0 FPs and 12 FNs. The fact that item classifications had changed again would suggest that the ILAP procedure could have been carried further. However, cost considerations and the fact that there were no FPs, although a number of FNs remained, made it seem unlikely that additional passes through this procedure would lead to further improvement in the identification of biased items.

The results for bias condition 4 showed that the ILAP procedure completely eliminated false positives, but it did not have as much of an impact on reducing false negatives. This could have resulted from items that were originally created to be only very weakly biased. To examine this possibility, average probabilities for Group B that responded correctly to each item from the unbiased test were compared to those for the biased test that contained the 36 biased items. It was found that all of the FNs were among the 16 most weakly biased items (i.e., those biased items with the smallest differences in average probabilities).

Mixed Test Bias Condition

There were 18 items biased against each group for bias condition 5 [$\theta_2 \sim N(-.5, 1.0)$ and $r(\theta_1,\theta_2) = .6$ for both groups]. When the M-DIL method was applied to this condition, two biased items were found to be unbiased for the initial linking. After the second iteration, however, the χ^2 value for one more biased item was statistically significant, and remained significant after convergence on the third iteration.

Bias condition 6 also had 18 items biased against each group, but it had different mean values of the θ_2 distribution for each group. In addition, the correlation between the focal and incidental dimensions was greater. Both the M-DIL and M-LTP procedures yielded the same results, except that two iterations were required for the M-DIL method to reach convergence, whereas four iterations were needed for the M-LTP method. In fact, after the third iteration for the M-LTP method, the same set of items was identified as biased, which also had been the case with the M-DIL method, but the fourth M-LTP iteration added one additional false negative.

Discussion

Non-iterative use of Divgi's linking method had essentially the same problems as the linking methods of both Warm (1978) and Linn, Levine, Hastings, and Wardrop (1981) in an earlier study (Lautenschlager & Park, 1988). Because Divgi's method is perhaps one of the best metric linking methods, considering sophistication, simplicity, accuracy, and cost, it is useful to know that iterative use of this method could effectively reduce the number of FP and FN identifications.

For bias condition 1, both the M-DIL and M-LTP linking methods were accurate in correctly classifying truly unbiased items. This was also true of the M-DIL method for bias condition 2. Subsequent iterations had little effect on the outcome. However, for unidirectional bias condition 3 (with 28 biased items), the M-LTP method proved clearly inferior, resulting in a very large number of false positives. The M-DIL method did overcome the FP problem after convergence, but a substantial number of FNs remained. In the presence of the most pervasive unidirectional bias (condition 4), both FPs and FNs remained numerous, even after convergence of the M-DIL method, and the initial M-LTP results again looked poor. These results imply that even very sophisticated metric linking methods have difficulty correctly classifying items as biased or unbiased as the number of biased items grows larger.

The mixed bias conditions (5 and 6) were generally less problematic for the correct identification of biased and unbiased items. Although the total number of biased items was great, the fact that bias was attributable to different incidental dimensions (θ_2) could have led to a canceling of effects attributable to item bias (Lautenschlager & Park, 1988). Furthermore, it is notable that the number of biased items was only a third of the total number of items within each group, and thus θ would be estimated from a preponderance of truly unbiased items.

The practical import of the findings in these mixed bias conditions must be tempered by at least two considerations. First, it is reasonable to question the likelihood that two completely different incidental dimensions would affect the test performance of two separate groups on a given test. Such conditions could be difficult to find in actual test situations. Second, it rarely will be known a priori whether a real test contains unidirectional or mixed directional forms of item bias.

The number of FNs in the unidirectional bias conditions appears to pose a potential problem. Drasgow (1987) asserted that it is meaningful to make the distinction between significant differences and practically important differences in IRT item bias analysis. He addressed this issue at the total test level, but it may also be examined at the item level. If weakly biased items were flagged as statistically significant, they would be labeled as biased in a statistical sense, but not necessarily in a practical sense. In the present study the most weakly biased items tended to have nonsignificant χ^2 values. It could be that these items were like unbiased items; however, one question remains unresolved.

When examining the results for bias condition 4, the χ^2 values for the unbiased items changed throughout the iterations. The weakly biased items tended to remain statistically unbiased throughout the iterations, though, and the χ^2 values remained largely unchanged. Also, 9 FPs remained at the convergence of the M-DIL method in bias condition 4, but the status of the FN items was virtually constant across the iterations. In addition, only the status of 5 out of the 16 FNs was changed, although the χ^2 values for all 9 FPs were changed to smaller nonsignificant values after the scale purifications described for the application of the ILAP procedure.

The issue of dealing with the FN problem is closely related to setting the α level for the χ^2 test. In examining the data for truly biased items in terms of probability difference between the items that were detected as unbiased and those items detected as biased, the items that had smaller than a .1 probability difference tended to have Lord asymptotic χ^2 values smaller than 13. Because this much probability difference looks trivial, the critical value for the χ^2 test could be set at slightly above 13—that is, at the .001 or .0005 α level to keep these items in the final version of the test. It is important to consider, however, the cumulative measurement bias for these weakly biased items. As stated above, these items tended to have smaller χ^2 values, regardless of the number of biased items in a test. Therefore, selecting a critical α level is an important issue (McLaughlin & Drasgow, 1987).

Another finding related to setting the α level is that moderately or strongly biased items almost always had large χ^2 values. This was true even when Warm's metric linking method had been used in a previous study non-iteratively for a simulation involving 46 biased items out of a 54-item test (Lautenschlager & Park, 1988). This could indicate that any biased items greatly influencing the total test score would be easily detectable using most metric linking methods. Judging from the results obtained in the present simulations, it was difficult to miss those items in IRT item bias analysis. Thus, if the possibility of strongly or moderately biased items after an item bias analysis is excluded, any remaining "weakly" biased items may not have much influence on the total test score.

Drasgow (1987; McLaughlin & Drasgow, 1987) reported that when estimated θ parameters were used in place of true θ values in simulation studies, Lord's χ^2 values tended to be inflated. Because estimation of person parameters was influenced by the presence of biased items in the item bias analyses, it was desirable here to examine whether Drasgow's results would be replicated, even though a different method was being used to simulate item bias. The χ^2 values for almost all items in the present study, too, tended to be larger when estimated θs were used than when true θ values were used in calculating the statistics. Therefore, using larger critical χ^2 values

would seem to be more appropriate.

The findings in the unidirectional bias conditions appear to be somewhat consistent with the recently reported results of Candell and Drasgow (1988), who used a different method for simulating item bias and a slightly different method for iterative linking. Both studies provided evidence that it is rare for truly unbiased items to be identified as biased when iterative linking is used, but false negative rates were generally higher in the present study. This could be attributed to the higher proportion of biased items involved, to the metric linking method employed, or to the differences in the amount and way in which item bias was simulated. Both studies also presented consistent evidence that non-iterative linking should be avoided in IRT bias analysis. Based on the results obtained here, the M-DIL method is preferable to the M-LTP method, because it generally converged in fewer steps with equivalent or better results.

Rather than using the M-DIL method alone, however, using a combination of iterative linking with θ scale purification led to more accurate classification of biased and unbiased items. The ILAP method was developed in an attempt to deal with the most pervasive unidirectional item bias condition simulated (condition 4). The results of this limited application of the ILAP procedure suggest that purification of θ scales, even after iterative parameter linking, could be a useful step in improving IRT item bias analysis.

Even a sophisticated way of linking parameter metrics cannot, however, exert its full potential if there are poor item parameter estimates. This fact further adds to the admonition of Lautenschlager and Park (1988), that the "common" items for linking θ metrics are those that are unbiased to begin with. The hope raised here is that the use of the ILAP method may well make it possible to better isolate these "common" items. Even though such complex procedures require more time, effort, and expense, they do appear to hold promise for resolving a significant problem in IRT item bias detection. Hopefully, feasible and efficient methods for disentangling biased from unbiased items can be further developed and refined so that they can be confidently implemented with actual test data.

References

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37–48.

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

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: CBS College Publishing.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

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.

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

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

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

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (19–29). Amsterdam: Swets and Zeitlinger.

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

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

McCauley, C. D., & Mendoza, J. L. (1985). A simula-

tion study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9,* 389–400.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11,* 161–173.

Park, D. G. (1988). *Investigations of item response theory item bias detection.* Unpublished doctoral dissertation, Department of Psychology, University of Georgia, Athens.

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

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

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

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Warm, T. A. (1978). *A primer of item response theory* (Technical Report No. 941078). Washington DC: U.S. Coast Guard Institute.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST: A user's guide.* Princeton NJ: Educational Testing Service.

Acknowledgments

This article is based in part on data from the first author's Doctoral Dissertation completed at the University of Georgia, U.S.A. Both authors contributed equally to this research. The authors thank two anonymous reviewers for their comments.

Author's Address

Send requests for reprints or further information to Dong-Gun Park, Department of Psychology, Ajou University, Suwon, Korea; or Gary J. Lautenschlager, Department of Psychology, The University of Georgia, Athens GA 30602, U.S.A.