IRT DIFFERENTIAL ITEM FUNCTIONING: AN EXAMINATION OF ABILITY SCALE PURIFICATIONS

GARY J. LAUTENSCHLAGER AND VICKI L. FLAHERTY
University of Georgia

DONG-GUN PARK
Korea University

Item Response Theory (IRT) differential item functioning (DIF) methods were employed to determine the accuracy of item classification as biased or unbiased. The procedure involved a combination of Drasgow's iterative item parameter linking method and Lord's ability scale purification procedure. Biased items were created using a two-dimensional noncompensatory IRT model. Previous research had demonstrated that the iterative linking and ability scale purification (ILAP) method held promise for discerning biased from unbiased items in one simulated unidirectional DIF condition. The present study found that ILAP was at times more effective than iterative linking alone primarily by reducing false negatives and occasionally reducing false positive misidentifications. The choice of significance level employed for detection of DIF did influence the convergence of the ILAP method. The method does hold some promise for improving correct classification of test items as biased or unbiased.

Evidence of the Usefulness of Ability Scale Purification

Employment and educational tests have frequently been subject to the criticism of being biased. However, the notion that entire tests are biased is not nearly so compelling as the belief that certain items on a given test are the sources of such bias (Berk, 1982). Although classical test theory methods for detecting differential item functioning (DIF) have been developed (Shepard, Camilli, & Williams, 1985), item response theory (IRT) has been

Correspondence concerning this article should be addressed to Gary J. Lautenschlager, Department of Psychology, University of Georgia, Athens, GA, 30602. Electronic mail may be sent to GARYLAUT@UGA.CC.UGA.EDU.

Educational and Psychological Measurement, Vol. 54 No. 1, Spring 1994 21-31

considered the theoretically preferred method for the detection of DIF owing to the sample invariant properties (Lord, 1980).

Biased items must necessarily tap secondary ability dimensions that are not reflected in responses to unbiased items (Crocker & Algina, 1986). It is these extraneous sources of variation that affect performance in a way that differs systematically for some subpopulations. Such factors may result in an unfair advantage to members of one subpopulation over others. Items on which differential performance is related to an additional unintended ability dimension might be termed biased items in a given context.

IRT methods for DIF detection are based on comparing item-characteristic curves (ICCs), which were estimated separately for each group. If a given item is unbiased, then the ICCs for that item, as estimated for the different groups, should be the same. When the estimated ICCs of the same item differ for two groups by more than sampling error, then DIF is suspected (Lord, 1980). Various methods have been proposed for examining differences in ICCs (McCauley & Mendoza, 1985; Shepard, Camilli, & Williams, 1985).

Issues in IRT Differential Item Functioning Analysis

Differential item function analysis in IRT is based on the belief that it is possible to put the ability and item parameters estimates obtained from samples drawn from the subpopulations onto a common scale. Various methods have been developed that could be employed to link the metrics to a common scale to enable comparison of the ICCs (Divgi, 1985; Linn, Levine, Hastings, & Wardrop, 1981; Stocking & Lord, 1983). However, it has been shown that this linking step has some potential problems. The most critical obstacle to overcome is the determination of which specific items to use in linking the metrics for the parameter estimates obtained from the separate groups. The paradox of IRT item bias analysis is that items that should be involved in linking ability (Θ) and item metrics are the truly unbiased items, the very items one hopes to identify through the item bias analysis itself (Lautenschlager & Park, 1988; Park & Lautenschlager, 1990).

Metric Linking and Ability Scale Purifications

A method for identifying DIF items developed by Park (1988; Park & Lautenschlager, 1990) combines a modified version of Drasgow's (1987) iterative linking method and Lord's (1980) ability purification procedure. This DIF detection method is referred to as iterative linking and ability scale purification (ILAP). The procedure is given as follows:

1. Item and person parameters are estimated separately for each group, placed on the same scale using Divgi's (1985) metric linking method, and item bias statistics are computed.

2. Item parameter estimates are relinked onto a common scale using only those items found to be unbiased in the previous step, and item bias statistics are recomputed for all items.
3. This linking process continues until the same set of items is found to be biased on two successive trials.
4. Person parameters are reestimated separately for each group using only those items flagged as unbiased from the final iteration of the previous step. (Ability Scale Purification)
5. New ability estimates are held constant, and item parameters are reestimated separately for each group for all items on the test.
6. Linking constants are estimated based only on those items flagged as unbiased preceding the purification step, and new item bias statistics are computed.
7. The linking process continues until the same set of items is classified as biased on two successive trials.
8. If the classification of items has changed since this purification step began, then return to Step 4. If no classification changes have occurred, then stop.

Park and Lautenschlager (1990) suggested that the ILAP procedure might be useful for improving IRT-DIF analysis. They found that the ILAP method helped to reduce the number of false positives but did not have as much impact on false negatives for one pervasive item bias condition. The purpose of the present study was to examine the effects of applying the ILAP method across a broader range of item bias conditions.

Method

The procedure adopted to simulate biased item responses was that developed by Park (1988; Lautenschlager & Park, 1988), where DIF items were essentially unidimensional for one group but multidimensional in the other (Hambleton & Swaminathan, 1985). This involved the use of an incidental ability dimension to influence performance on biased items. This definition of item bias involves a focal dimension (Θ_1) and a second incidental dimension (Θ_2), which can influence performance on some items.

The unidimensional three-parameter logistic IRT model was used to generate data for all unbiased items. The two-dimensional version of the multidimensional IRT model proposed by Sympson (1978) was employed to generate item response data for each biased item (Lautenschlager & Park, 1988). Sympson's (1978) multidimensional noncompensatory model is given as

$$P_{ij}(\Theta_{ih}) = c_j + (1 - c_j)\,(\text{II}\,\{1 + \exp[-1.7a_{jh}(\Theta_{ih} - b_{jh})]\})^{-1}$$

where Θ_{ih} is the ability parameter for person i for dimension h, a_{jh} is the discrimination parameter for item j for dimension h, b_{jh} is the difficulty parameter for item j for dimension h, and c_j is the pseudo-guessing parameter for item j.

Ansley and Forsyth (1985) provided a justification for using Sympson's noncompensatory model over the other models because the noncompensatory view of dimensionality is more reasonable for most well-constructed achievement tests. In addition, they reported that this model produced data with properties very similar to those of actual achievement test data.

The ability dimensions employed is referred to as Θ_1, or the focal dimension common to all items and groups, and an incidental dimension, Θ_2, producing item bias in a subset of items for members of one group. 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 from .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 about .70. Item discrimination values for Θ_2 were centered at .50, with a standard deviation of about .10. The c_i parameters were set at .20 for all items.

Data Set Characteristics

Sets of data were generated to simulate item responses to multiple-choice items having four response options. Only two groups of simulated examinees were employed, hereafter referred to as Group A and Group B. Each data set reflected the responses of 1,000 examinees to 54 items. The number of DIF items in a given simulated test was fixed at either 18, 28, or 36 biased items of the 54 items on the test.

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 assuming a common c_i value was used to generate item response data on all items for Group A examinees and for all unbiased items for Group B examinees. The generation of item responses for the DIF items involved the use of the two-dimensional version of Sympson's (1978) model. Two types of normal ability distributions on Θ_2 were generated for the B group examinees by 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 set to either .60 or .90 in the population.

Table 1 presents a description of the data sets created based on combinations of the number of DIF items, trait score distributions on the focal and incidental ability dimensions, and the population correlation of these trait dimensions. Each data set involved simulation of bias directed against Group B.

Table 1
Characteristics of Data Sets for DIF Bias Conditions

Bias condition	Group	Number of unbiased items	Number of biased items	Normal Θ_2 distribution	
				M	*SD*
1	A	54	0	—	—
	B	36	18	0.0	1.0
2	A	54	0	—	—
	B	36	18	–.5	1.0
3	A	54	0	—	—
	B	26	28	–.5	1.0
4	A	54	0	—	—
	B	18	36	–.5	1.0

Note. For both the A and B groups, 1,000 examinees were simulated. Unbiased items were created using the unidimensional three-parameter logistic model. Biased items were created for Group B using Sympson's (1978) multidimensional model. Θ_2 is the incidental trait. Each condition was simulated with the population correlation between the focal and incidental trait set at $r = .6$ and again at $r = .9$.

Analysis

The LOGIST computer program was used to estimate the item and person parameters (Wingersky, Barton, & Lord, 1982). Lord's (1980) chi-squared item bias statistic was used to indicate the potential for item bias. A significance level of $\alpha = .005$ was used for indicating "detected" DIF bias. For a subset of these analyses, $\alpha = .001$ was employed to examine how the choice of a significance level affected the results.

The linking of IRT item parameters was accomplished by using a modification of Drasgow's (1987) iterative linking procedure. Divgi's (1985) minimum chi-square method for linking was substituted in place of the more complex Stocking and Lord (1983) method used by Drasgow (1987). Relinking of parameters was accomplished by applying Divgi's method to subsets of items that had been flagged as unbiased in the immediately preceding iteration to determine new linking constants. Item bias statistics were then recalculated to determine whether any item classifications had changed from the previous iteration.

Ability scale purification was conducted after convergence was achieved by using the iterative linking method described earlier. Each purification step involved using another application of the LOGIST program, where only those items classified as unbiased were selected and employed to generate "purified" ability estimates (i.e., ability estimates that are not influenced by items that had been identified as biased after convergence of the iterative linking step). These new ability estimates were then held fixed in a subsequent

LOGIST analysis and item parameter estimates were obtained for all items. At this point, iterative linking was again employed to determine whether classification of items as biased or unbiased had changed. Any change in item classifications indicated the need for an additional purification step. Convergence of the purification procedure occurred either when no changes were found in the classification of items compared to the results of the previous step or when the procedure reached a recursive loop.

A procedure termed "unbiased purification" was employed to provide another baseline for the accuracy of biased item detection. This approach involved selecting only those items that had been created at the outset to be unbiased and using them to estimate ability parameters. These new ability estimates were subsequently employed in a single purification step. Item bias statistics were then obtained and the outcome served as another criterion.

The logic in this procedure was that these ability estimates should be the "best" because they are based on only those items created using the unidimensional IRT model. However, it is also important to note that the use of only the unbiased items to estimate ability does reduce the length of the test and, in turn, may reduce the precision of those estimates. Hence the "best" ability estimates in the sense intended here may not always be good ones, especially if test length decreases considerably. It would appear that this method would provide a useful comparison baseline for the accuracy of item classification. It was expected that the identification of DIF under conditions blind to the true status of the items would likely prove no better than this unbiased purification baseline.

Results

The outcome regarding the number of false positives (FPs) and false negatives (FNs) at convergence of the initial iterative linking procedure and at the end of each purification step are presented in Table 2. In addition, the results of the true purification analyses are set forth in the last row of this table.

In general, the higher the correlation between the two ability dimensions, the greater the likelihood of false negative misidentifications at the end of the first iterative linking phase. False positives were only a problem when $r_{\Theta_1 \Theta_2} = .6$ for the most pervasive item bias condition. It should be noted that this would be the stopping point for identifying DIF items if one were following the iterative linking procedure of Drasgow (1987; Candell & Drasgow, 1988).

The first purification step was generally effective in reducing the number of false negative misidentifications when $r_{\Theta_1 \Theta_2} = .6$. It is interesting to note that additional FPs were eliminated in some conditions after convergence of the initial iterative linking process. In the most pervasive bias condition, the

Table 2

Iteration History of False Positives and False Negatives for Initial Estimation and at Each Purification Step ($\alpha = .005$)

		DIF bias condition															
		$r_{\Theta_1\Theta_2} = .6$								$r_{\Theta_1\Theta_2} = .9$							
Purification		1[a]		2[a]		3[a]		4		1[a]		2		3[b]		4[a]	
Step	Iteration	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
0	—	1	3	1	3	0	7	9	16	0	8	0	7	1	6	1	15
1	1	0	2	0	1	1	6	2	15	1	8	0	7	1	6	1	16
	2	0	2	1	1	1	6	1	15	1	8	0	7	1	6	0	16
	3			1	1												
2	1	1	3	1	3	1	7	1	11	0	8					1	16
	2	1	3	1	3	1	7	0	11	0	8					1	16
	3							0	11								
3	1	0	2	1	1	1	6	1	10	1	8					0	16
	2	0	2	1	1	1	6	1	10	1	8					0	16
4	1							0	10								
	2							0	10								
5	1							0	10								
	2							0	10								
Unbiased purification		2	3	2	3	1	6	1	8	1	7	0	7	0	6	0	6

Note. Purification "Step 0" is the result at convergence of the iterative linking process for the initial estimation phase before any purification of the ability estimates was performed.

a. The purification process for this condition failed to converge.

b. For this condition, the iterative linking procedure at initial estimation failed to converge.

Table 3
Iteration History of False Positives and False Negatives for Initial Estimation and at Each Purification Step (α = .001)

Purification		DIF bias condition $r_{\Theta_1\Theta_2} = .6$							
		1		2[a]		3		4	
Step	Iteration	FP	FN	FP	FN	FP	FN	FP	FN
0	—	0	3	0	3	0	10	9	17
1	1	0	3	0	3	0	9	2	17
	2	0	3	0	3	0	8	0	16
	3					0	8	0	16
2	1			0	4	0	8	1	14
	2			0	4	0	8	0	13
	3							0	13
3	1			0	2			0	14
	2			0	2			0	14
4	1			0	5			0	13
	2			0	5			0	13
5	1			0	3			0	13
	2			0	2			0	13
	3			0	2				
Unbiased purification		0	5	0	5	0	7	0	11

Note. Purification "Step 0" is the result at convergence of the iterative linking process for the initial estimation phase before any purification of the ability estimates was performed.
a. The purification process for this condition failed to converge.

purification had a dramatic impact on reducing the number of FP misidentifications. Purification also reduced the number of FNs.

For the bias conditions where $r_{\Theta_1\Theta_2} = .9$, false positives only appeared at the end of the initial iterative linking step (Step 0) in the two conditions with the largest number of DIF items. There were considerably more FN item misclassifications in the two 18 biased item conditions than when $r_{\Theta_1\Theta_2} = .6$. The first purification resulted in adding an FN in Bias Condition 1 and in removing an FP and adding an FN in Bias Condition 4. In sum, the first purification had little impact on any of these bias conditions.

It should be noted that when $r_{\Theta_1\Theta_2} = .9$ for Bias Condition 3 that the initial iterative linking process also failed to converge. In effect, a particular FP item appeared at linking Iteration 4 and then disappeared on Iteration 5 only to reappear at Iteration 6. As no other items changed classification during this process, a conservative rule was invoked to consider that item flagged as biased at this point. (Additional consideration of these outcomes appears in the Discussion section.)

Comparing results of the final purification step with the true purifications indicated that this baseline was achievable through the application of the ILAP procedure. In fact, it was possible to do even slightly better than this baseline in some instances.

Complications arose in attempting to use the purification process in subsequent purification steps. The purification process converged in just three of the eight DIF conditions. Two of the converged sets of results were no different from the results obtained at the end of the initial iterative linking step (i.e., before any purification was attempted).

To determine whether this failure to achieve a clear convergence was affected by the use of too liberal a significance level in the chi-squared tests (McLaughlin & Drasgow, 1987), an examination of the ILAP procedure was made employing $\alpha = .001$ for the $r_{\Theta_1 \Theta_2} = .6$ item bias conditions. The results based on this more restrictive criterion for flagging items as biased are presented in Table 3. Convergence of the purifications was obtained in all but one condition. As might be expected, false positives were reduced or eliminated, and FNs increased by use of a more stringent significance level.

Discussion

Consistent with the findings of Park and Lautenschlager (1990), the use of ability scale purifications in conjunction with iterative item parameter linking (Drasgow, 1987) can improve the correct detection of DIF items. The improvement was greatest in the present study when the correlation between the focal and incidental traits was $r_{\Theta_1 \Theta_2} = .6$. Related to this issue it is interesting to note that Donlon (1984) reports a correlation of $r = .42$ between the Verbal and Quantitative sections scores on the Scholastic Aptitude Test. Due to the effects of attenuation, this sample correlation likely underestimated the correlation of the two focal latent traits measured by those tests. The combination of these traits, as jointly relevant to performance on a subset of test items for a test measuring a quantitative ability, could lead to the detection of DIF items if the subgroups differed on the incidental reading dimension (Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984).

The failure of the purification procedure to reach convergence in numerous cases may seem problematic. However, even when convergence was not achieved, the results still tended to shift toward better recovery of true item status at the close of one purification step before shifting back in the direction of the results obtained at the end of the initial linking step (Step O). This outcome was true even in the most pervasive bias conditions. This failure to achieve convergence may not be a limiting problem for the method. In fact, as noted previously, the iterative parameter linking method itself was not

immune to this convergence problem. The ILAP method would appear to have some potential for better isolating subsets of potentially biased items.

In practice, the failure of the ILAP procedure to converge may require the consideration of all items flagged as biased in the course of the loop as potentially biased. As in any investigation of this issue, such items should be inspected to discern logical bases that support their removal or modification, as has been advocated by Berk (1982), among others.

The choice of the alpha level to use in detecting potentially biased items merits consideration. Use of $\alpha = .005$ resulted in more failures of the purification process to reach convergence than when a more restrictive alpha level was used. This problem may largely be due to the detection of FPs, which reduces the pool of items used to obtain ability estimates during a purification step. This problem would appear to be readily resolved by adopting a smaller alpha level; however, this step will result in a trade-off by increasing the FN rate. Clearly, the consequences of such a trade-off would need to be considered in practice.

The failure to detect items that contain bias may lead to bias in overall test scores. Drasgow (1987) has indicated that after iterative linking alone the cumulative bias present in the remaining item pool may be inconsequential. Certainly, the trade-offs in terms of the extra costs involved with implementation of the ILAP method may need to be weighed against the benefits of greater confidence in the discrimination of biased from unbiased items. Iterative linking alone is not very expensive, but the purification steps require additional estimation using an IRT parameter estimation program.

Although the FP problem is not great, it is interesting to note that the ILAP method does further reduce this problem. As Berk (1982) and others have noted, it is unwise to discard items solely on the basis of a statistic without inspecting the item for the possible causes of bias.

The correct classification of items as biased and unbiased is critical to the construction of "fair" tests from the standpoint of measurement equivalence (Drasgow, 1984). Any procedure that can improve the discrimination of biased from unbiased items stands to make such a goal more attainable. The ILAP method holds promise for achieving more effectively that goal.

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, MD: Johns Hopkins University 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 a common metric in item response theory. *Applied Psychological Measurement, 9*, 413-415.

Donlon, T. F. (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

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 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 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.

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*, 161-174.

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

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

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 22*, 77-105.

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* (Laboratory of Psychometric and Evaluation Research Rep. No. 90). Amherst: University of Massachusetts, School of Education.

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: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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