A Simulation Study of Item Bias Using a Two-Parameter Item Response Model

Cynthia D. McCauley
Center for Creative Leadership

Jorge Mendoza
Texas A & M University

Possible underlying causes of item bias were examined using a simulation procedure. Data sets were generated to conform to specified factor structures and mean factor scores. Comparisons between the item parameters of various data sets were made with one data set representing the "majority" group and another data set representing the "minority" group. Results indicated that items that required a secondary ability, on which two groups differed in mean level, were generally more biased than those items that do not require a secondary ability. Items with different factor structures in two groups were not consistently identified as more biased than those having similar factor structures. A substantial amount of agreement was found among the bias indices used in the study.

Bias in tests has become an important issue in recent years. In assessing bias in a test, a criterion that is considered less biased is used. One approach, generally used in selection research, has been to use a criterion external to the test, such as job performance, in assessing test bias (e.g., Gael, Grant, & Richie, 1975). The present study focused on a second approach to the study of bias: identification of bias in the absence of an external criterion. Drasgow (1984) has referred to this conceptualization of bias as "measurement equivalence": test scores are unbiased if they have identical relations with the attribute measured by the test in all subgroups of interest. Measurement equivalence is investigated by examining bias in individual items from the test of interest. An item is considered biased if individuals with equal ability, but from different groups, do not have the same probability of answering the item correctly (Ironson, 1982; Shepherd, Camilli, & Averill, 1981).

Item response theory (IRT) provides the most theoretically sound framework for studying item bias (Lord, 1980; Wright, 1977). The sample invariant quality of the parameters of IRT models makes this method of item bias detection least sensitive to distributional differences in groups. According to IRT, the probability of answering an item correctly, $P(\theta)$, is a function of ability (θ) and the three item parameters, the discrimination parameter, a, the difficulty parameter, b, and the guessing parameter, c. For an item to be unbiased, it must yield equal $P(\theta)$ for examinees of equal ability regardless of group membership. In other words, the item characteristic curves (ICCs) computed separately on two groups must be the same.

Multidimensionality

Item bias has been defined as differential probability of answering an item correctly when having the same ability but belonging to different subgroups. The question arises concerning what causes this differential probability of answering an item correctly. Two suggestions have been made to answer this question; both deal with multidimensionality.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 4, December 1985, pp. 389–400

0146-6216/85/040389-12$1.85

One possible underlying cause of item bias may be "multidimensionality confounding differences on a primary trait with differences on a secondary trait" (Linn, Levine, Hastings, & Wardrop, 1981, p. 161). If an item measures the exact same ability in two groups and only that ability, then this differential probability of a correct response when having the same ability would not occur. However, answering the item may require a secondary ability as well as the primary ability (the ability the test was intended to measure). If the distributions of the secondary trait are different in the two groups, bias may result. For example, some items on a test of numerical ability may require reading comprehension. One group may on average have lower reading comprehension ability. Requiring reading comprehension on these items may result in a lower probability of answering the items correctly for members of this group than would be expected given their numerical ability. Verbal math problems of this type have been found to be a source of bias against blacks (Shepherd, Camilli, & Williams, 1984).

Multidimensionality may lead to item bias in an additional manner. An item may require an additional ability in one group or may require different secondary abilities in the groups. In other words, different test factor structures may exist in the two groups. Item bias in this situation would occur when an item "systematically measures one thing for one group and a different thing for a different group" (Devine & Raju, 1982).

The assumptions of IRT require a test to be unidimensional. If the presence of bias is due in some way to multidimensionality, then this would explain why a biased item does not fit the IRT model. Robustness studies (Drasgow & Parsons, 1983; Reckase, 1979) have found that parameter estimation is still possible with moderate amounts of multidimensionality. The present study stayed within the limits set by robustness studies in investigating item bias.

Present Study

The present study investigated several unanswered questions concerning item bias in the IRT framework. Data were simulated under the assumption that a secondary ability was required on some of the items on an otherwise unidimensional test. The number of items that required the secondary ability and the mean difference between two groups on the secondary ability were the main variables manipulated. Biased items in the absence of mean differences between the two groups on the primary ability was chosen as the simplest case to investigate. According to IRT, mean group differences on the primary ability measured by a test should not lead to bias; but possible interaction effects from the combination of mean group differences on the primary and secondary abilities were not investigated in the present study.

Past monte carlo studies of item bias have generated biased items by selecting different values of *a* and *b* parameters for two groups on each item (Rudner, Getson, & Knight, 1980). This method of generating bias does not help in understanding what may cause bias in items. Since multidimensionality has been suggested as related to item bias, the present study used this concept in generating biased items. Two questions were addressed:

1. Will item bias indices identify test items that require a secondary ability, on which two groups differ in mean level, as more biased than those items that do not require a secondary ability?
2. Will item bias indices identify items that have different structures in different groups as more biased than those items having similar factor structures?

Another issue addressed by the study concerned the comparability of the bias indices. Several studies have compared item bias indices (Shepherd et al., 1981, 1984). However, the simulation procedure allowed the study of agreement of bias indices under varying conditions.

Method

Data Generation

The first step in the data generation process was to generate propensity scores, based on the factor analysis model, which were later to be transformed

into (1/0) item scores. The process for each of the 20 data sets used was based on the equation

$$\mathbf{X} = \mathbf{ZF} + \mathbf{E} \quad , \tag{1}$$

where **X** is the resulting propensity matrix representing 1,000 subjects and 50 items;
Z is a computer-generated matrix containing normally distributed factor scores;
F is a specified factor structure matrix; and
E is a matrix of normally distributed error.

Each factor structure consisted of one general factor and 0 to 2 secondary factors. The data sets with one secondary factor had either 8 or 15 items loading on the second factor (the same 8 or 15 in each data set). The data set with two secondary factors had 8 items loading on the second factor and 7 items loading on the third factor. The loadings on the first factor were the same for all data sets and ranged between .55 and .75. Loadings on the second factor(s) ranged between .35 and .50. The mean of the factor scores on the first factor was zero, while the means on the secondary factor(s) were 0, $-.5$, or -1.

The item propensity scores were transformed to 1/0 responses by a threshold method (Drasgow & Parsons, 1983). Item propensity scores equal to or above the item's threshold (randomly chosen from a normal distribution of z scores) were coded as 1 (correct); scores below threshold were coded as 0 (incorrect). Several of the 1/0 data sets created were subjected to confirmatory factor analysis (Jöreskog & Sörbom, 1979), and in each case the intended factor structure was recaptured.

Parameter Estimation

The LOGIST computer program (Wingersky, 1983; Wingersky, Barton, & Lord, 1982) was used to obtain maximum likelihood estimates of item parameters and person abilities for each data set under the two-parameter logistic IRT model. The two-parameter model ($c = 0$) was chosen over the full three-parameter model for practical reasons; the c parameter is difficult to estimate, convergence problems are minimized, and computer time is reduced. The indices developed with the three-parameter model are applicable to the two-parameter model.

Generating Bias

Comparisons between the item parameters of various data sets were made with one data set representing the "majority" group and another data set representing the "minority" group. These conditions are presented in Table 1. The first type of comparison, data sets with identical factor structures and equal mean factor scores, provided a baseline for the bias indices in other comparisons. The comparisons of primary interest were those involving data sets with unequal mean factor scores or different factor structures. The question of interest in each comparison concerned whether items theoretically created to be "biased" (either because the item required a secondary ability on which one group scored at a lower mean level or because the item loaded differently in the two groups) would be identified as more biased than those items not created to be biased. Since the items were calibrated separately for each group, the a and b values for the minority group were put on the metric of the majority group by using the equating formulas from Warm (1978).

Bias Indices

Area indices. Areas were approximated following the procedure outlined by Linn, Levine, Hastings, and Wardrop (1980). Distances between $\theta = -3$ and $\theta = +3$ were divided into 600 intervals. The distance between the ICCs at the midpoint of the interval was multiplied by the width of the interval (.01).

Let P_{j1} and P_{j2} be the height of ICCs of the base (majority) and comparison (minority) groups respectively evaluated at the midpoint of the jth interval. The distance between ICCs at the midpoint was defined as $D_j = |P_{j1} - P_{j2}|$. Let $f = 1$ if $P_{j1} > P_{j2}$ and $f = 0$ if $P_{j1} < P_{j2}$. The following area indices were used:

1. Base high area: $I_1 = (.01)\sum fD_j$.
2. Base low area: $I_2 = (.01)\sum (1 - f)D_j$.
3. Unsigned area: $I_3 = I_1 + I_2$.
4. Signed area: $I_4 = I_1 - I_2$.
5. Square root of sum of squares: $I_5 = [(.01)\sum D_j^2]^{1/2}$.

Table 1
Summary of Data Set Comparisons

Factor Structures and Conditions	"Majority" Group				"Minority" Group			
	Number of Items Loading on Second Factor(s)	Mean Factor Scores			Number of Items Loading on Second Factor(s)	Mean Factor Scores		
		M_1	M_2	M_3		M_1	M_2	M3
Identical factor structures--equal mean factor scores								
1	8	0	0	-	8	0	0	-
2	15	0	0	-	15	0	0	-
3	0	0	-	-	0	0	-	-
Identical factor structures--unequal mean secondary factor scores								
4	8	0	0	-	8	0	-1	-
5	8	0	0	-	8	0	-.5	-
6	15	0	0	-	15	0	-1	-
7	15	0	0	-	15	0	-.5	-
Different factor structures--equal mean factor scores								
8	0	0	-	-	8	0	0	-
9	0	0	-	-	15	0	0	-
10	15	0	0	-	15	0	0	0

Parameter difference measures.

1. Lord's (1980) significance test: Given Groups 1 and 2, for item i,
   $I_6 = \chi^2$
   $= (\mathbf{X}_{i2} - \mathbf{X}_{i2})'(\mathbf{S}_{i1} + \mathbf{S}_{i2})^{-1}(\mathbf{X}_{i1} - \mathbf{X}_{i2})$,
   where $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ are vectors of a and b parameters for Groups 1 and 2; $\mathbf{S}_{i1}$ and $\mathbf{S}_{i2}$ are variance/covariance matrices for a and b in Groups 1 and 2. The chi-square value obtained is compared to the critical value with 2 degrees of freedom.
2. Difference in a parameters: $I_7 = a_1 - a_2$.
3. Difference in b parameters: $I_8 = b_1 - b_2$.

Analysis

Several methods were used to assess whether items created to be biased were identified as more biased by bias indices than those not created to be biased. First, within each condition, the mean item bias index of the biased items was compared to the mean index for the unbiased items. Also, the biased item mean index was compared with the mean index of those same items in a condition in which no bias was created (i.e., Conditions 1, 2, and 3).

Another approach was to see if individual items created to be biased were identified as biased using criteria currently in use.

1. An item was considered biased if the chi-square index (I_6) was significant ($\alpha = .01$; Lord, 1980). This is the only index with an associated significance test.
2. Bias indices were calculated for items in a comparison between two subgroups of the majority group. Then, the same index was calculated for the same items in a majority/minority comparison. When the index in the majority/minority comparison was substantially larger than the index in the majority/majority comparison, the item was considered biased (Angoff, 1982). The indices for the "majority/majority" comparisons were the baseline condition with the same factor structure as the majority/minority comparison. For example, Condition 5 is a majority/minority comparison; the appropriate majority/majority or baseline condition to contrast it with is Condition 1. When an index for an item in Condition 5 varied more than two standard deviations (SD) from the index for that item in Condition 1, then that item was considered biased using that particular index.

Results

Identical Factor Structures: Equal Factor Score Means

In Conditions 1, 2, and 3 the data sets had the same factor structure and equal factor score means. No bias was expected in these baseline conditions. The difference between the means of items loading and not loading on the secondary factor was evaluated with a t test in Conditions 1 and 2. None of the t values were significant at the .05 level.

Identical Factor Structures: Unequal Factor Score Means

In each case of these comparisons, the mean secondary factor score for the minority group was either .5 SD or 1 SD below the mean for the majority group. The mean of the biased items was compared to the mean of the unbiased items for each index on each comparison. The biased items were the items loading on the secondary factor with the mean factor score in the minority group being less than that in the majority group. The results of the t tests are reported in Table 2. All differences were significant except for I_7 (differences in a parameters). The significant mean differences were in the direction of more bias in the items created to be biased. These items had larger base high areas (I_1), smaller base low areas (I_2), larger unsigned indices (I_3 and I_5), larger signed area indices (I_4), larger chi-square values (I_6), and more negative differences in b parameters (I_8). As expected, the mean differences were generally larger in the conditions involving minority groups with second factor mean $= -1$ than in conditions involving groups with second factor mean $= -.5$. The differences in the ICCs of the items identified as biased were due primarily to differences in b parameters.

The mean of each index over the biased items was also compared to the mean in the baseline

Table 2
Mean Difference Between Biased and Unbiased Items in Conditions 4, 5, 6, and 7

Bias Index	Condition 4			Condition 5			Condition 6			Condition 7		
	M	(SD)[a]	t[b]	M	(SD)	t	M	(SD)	t	M	(SD)	t
I_1	.41	(.11)	9.83**	.24	(.06)	13.20**	.34	(.14)	8.53**	.20	(.11)	6.50**
	.01	(.02)		.02	(.04)		.01	(.04)		.02	(.03)	
I_2	.01	(.02)	-7.45**	.00	(.01)	-9.20**	.00	(.01)	-10.81**	.01	(.03)	-6.51**
	.09	(.06)		.06	(.04)		.19	(.10)		.10	(.07)	
I_3	.41	(.10)	8.55**	.24	(.06)	8.01**	.34	(.14)	4.00**	.21	(.10)	4.22**
	.10	(.06)		.08	(.05)		.21	(.09)		.11	(.06)	
I_4	.40	(.13)	10.34**	.24	(.06)	11.16**	.33	(.15)	12.67**	.19	(.12)	9.40**
	-.08	(.07)		-.04	(.07)		-.18	(.12)		-.08	(.08)	
I_5	.25	(.05)	11.79**	.15	(.03)	7.61**	.22	(.08)	3.79**	.14	(.09)	2.88*
	.07	(.04)		.05	(.03)		.13	(.06)		.07	(.05)	
I_6	51.95	(22.30)	6.06**	17.80	(8.70)	4.93**	40.00	(28.20)	3.84**	14.24	(9.40)	4.30**
	4.03	(4.03)		2.54	(2.10)		11.63	(7.30)		3.59	(3.20)	
I_7	-.27	(.30)	-1.60	-.16	(.12)	-1.42	-.07	(.25)	-.55	-.03	(.31)	.02
	-.10	(.14)		.08	(.17)		-.03	(.25)		-.03	(.19)	
I_8	-.43	(.16)	-8.82**	-.24	(.06)	-10.65**	-.35	(.17)	-11.82**	-.24	(.26)	-4.68**
	.08	(.08)		.04	(.07)		.18	(.14)		.08	(.09)	

[a] First row M is for biased items; second row M for unbiased items.
[b] If variances were significantly different, an approximate t statistic based on unequal variances was computed.
* $p < .05$
** $p < .01$

condition in which no bias was created. The results of this analysis closely approximated those found with the previous t tests: those items created to be biased were, on average, consistently identified as more biased than those items not created to be biased with all bias indices except I_7.

The number of false positives (items not biased but identified as biased) and false negatives (items biased but not identified as biased) for each bias index are presented in Table 3. Table 3 shows that as the amount of bias increases (i.e., the mean factor scores of the minority group goes from $-.5$ to -1), the number of false negatives decreases. I_2 and I_7 were the poorest at identifying the biased items and I_1 and I_6 were best. The chi-square index performed less accurately as the number of biased items on the test increased. There was a great deal of overlap in items identified as false positives by the different indices.

Different Factor Structures: Equal Factor Score Means

The third type of comparison involved groups in which some items differed in factor structure. In Conditions 8 and 9, the items in the majority group were unidimensional, while some items loaded on a secondary factor in the minority groups. Condition 8 involved 8 such items and Condition 9 involved 15. In Condition 10, the same 15 items loaded on a secondary factor in both majority and minority groups; however, in the minority group two secondary factors were involved.

The same type of analyses were performed on these data as described above in an attempt to determine if the items created to be biased were detected as more biased by the various indices. In this instance, however, "created to be biased" refers to items with different factor structures in the two groups.

Mean difference comparisons were made between bias indices of biased items and unbiased items in each of the conditions. Table 4 shows that results were not consistent across conditions. In Condition 8, there were significant mean differences for I_2, I_3, I_4, and I_6. The difference in means for I_4 was not in the expected direction (biased mean = $-.05$, unbiased mean = .01). In Condition 9, significant mean differences were found for I_1, I_2, I_4, and I_8. The differences were not in the expected direction for I_1, (biased mean = .02, nonbiased mean = .05), I_4 (biased mean = $-.05$, nonbiased mean = .02), and I_8 (biased mean = .07, nonbiased mean = $-.02$). There were no significant mean differences with Condition 10.

Table 3
Number of False Positives (FP) and False Negatives (FN) Identified in Conditions 4, 5, 6, and 7

Criterion and Bias Index	Condition 4[a]		Condition 5[a]		Condition 6[b]		Condition 7[b]	
	FP	FN	FP	FN	FP	FN	FP	FN
Two SD								
I_1	0	0	1	0	0	2	0	5
I_2	14	6	4	6	22	12	5	14
I_3	4	0	7	2	11	4	2	7
I_4	5	0	4	2	10	5	3	6
I_5	5	0	1	2	9	4	1	7
I_6	10	0	8	1	24	2	3	4
I_7	4	7	3	8	0	14	1	14
I_8	4	0	2	1	6	7	2	7
Chi-square								
I_6	2	1	0	2	18	2	3	6

[a] Total possible FN = 8; total possible FP = 42.
[b] Total possible FN = 15; total possible FP = 35.

Table 4
Mean Difference Between Biased and Unbiased Items in Conditions 8, 9, and 10

Bias Index	Condition 8 M	Condition 8 (SD)[a]	Condition 8 t[b]	Condition 9 M	Condition 9 (SD)	Condition 9 t	Condition 10 M	Condition 10 (SD)	Condition 10 t
I_1	.05	(.04)	.22	.02	(.03)	-2.90**	.03	(.03)	-.86
	.04	(.05)		.05	(.06)		.05	(.06)	
I_2	.10	(.08)	2.39*	.07	(.06)	2.14*	.07	(.06)	.65
	.03	(.04)		.03	(.04)		.05	(.06)	
I_3	.15	(.07)	3.79**	.09	(.06)	.29	.09	(.04)	-.18
	.07	(.05)		.08	(.05)		.10	(.06)	
I_4	-.05	(.11)	-2.16*	-.05	(.08)	-2.91**	-.03	(.09)	-.89
	.01	(.07)		.02	(.08)		.00	(.09)	
I_5	.10	(.07)	1.86	.06	(.04)	.31	.06	(.02)	-.49
	.05	(.04)		.05	(.03)		.07	(.04)	
I_6	4.52	(2.20)	3.58**	2.68	(2.60)	.43	2.78	(1.70)	.11
	1.76	(2.00)		2.37	(2.30)		2.68	(2.10)	
I_7	.15	(.27)	1.65	.08	(.14)	1.29	.09	(.09)	.76
	-.01	(.13)		.02	(.17)		.02	(.23)	
I_8	.10	(.24)	1.38	.07	(.13)	3.00**	.03	(.09)	.85
	-.02	(.08)		-.02	(.09)		.00	(.10)	

[a] First row M is for biased items; second row M for unbiased items.
[b] If variances were significantly different, an approximate t statistic based on unequal variances was computed.
* p < .05
** p < .01

The differences between biased items and the items in a baseline condition were also examined. The baseline for each of these conditions was the condition in which two groups were compared, both having the same factor structure as the majority group. The only index that showed a significant mean difference was I_7: there was a larger difference in the *a* parameters of biased items in the majority/minority comparison than in the baseline comparison for Conditions 8, 9, and 10 (Condition 8, $t_{14} = 2.55$, $p < .05$; Condition 9, $t_{28} = 1.91$, $p < .05$; Condition 10, $t_{12} = 2.00$, $p < .05$). The items were more discriminating (higher *a*s) in the majority group of the majority/minority comparison.

An examination was also made of the items identified as biased by each index using the two criteria. There was little variation in results across bias indices or criteria. Few false positives and many false negatives were identified. In other words, not many items were identified as biased using any index or criterion.

Agreement Among Indices

Agreement among bias indices was assessed by intercorrelating the indices in each condition. The correlation matrices were grouped according to type of comparison to see if this variable would affect the interrelations of indices. Since the correlation matrices within types of comparisons were highly similar, a representative of each type is presented. Table 5 presents the intercorrelations for Condition 1, in which no bias was created, Condition 4, in which subgroups had different mean secondary factor scores, and Condition 8, in which subgroups had different factor structures. There was generally high linear relationships among the area indices (I_1 through I_5). The exception was the low correlations of I_4 (signed index) with I_3 (unsigned index) and I_5 (sum of squares). The correlations between signed indices and unsigned indices were low because the relationships were nonlinear. With a signed index, a biased item would have a large positive or negative value, whereas an unsigned index would be

only positive. Therefore, the correlations between the absolute values of the signed indices and unsigned indices are presented in parentheses in the table. These correlations revealed a much stronger agreement among signed and unsigned area indices. However, in Condition 4 the original correlations indicated agreement among signed and unsigned indices, indicating that the bias created in this condition was generally one-directional (i.e., against the minority group).

Generally the correlations among the parameter difference indices (I_6 through I_8) were low. However, in Condition 4, I_8 was strongly related to I_6. This supports the notion that bias created by lowering the mean secondary ability of one group affects the *b* parameter. Across conditions, the greatest amount of agreement was generally found between I_3 and I_5. High negative correlations between I_4 and I_8 were also found, indicating that the biased items were easier for the majority group.

Discussion

A major purpose of the present study was to investigate underlying causes of item bias. One suggested cause was success on the item requiring

Table 5
Intercorrelations of Bias Indices in Conditions 1, 4, and 8

		I_2	I_3	I_4	I_5	I_6	I_7	I_8
I_1								
	1	-36	58	83	54	34	-18	-77
	4	-51	91	95	87	87	-39	-94
	8	-35	52	80	40	54	-06	-62
I_2								
	1		54	-82	56	50	-18	82
	4		09	-73	-04	-29	02	72
	8		62	-83	67	34	-04	88
I_3								
	1			03(89)	98	74	-32	02
	4			74(92)	98	87	-23	-73
	8			-09(82)	95	77	-09	28
I_4								
	1				00(92)	-08(57)	00	-96
	4				69(94)	79(63)	-17	-98
	8				-19(89)	10(47)	00	-93
I_5								
	1					**65**	**41**	09
	4					**83**	-25	-70
	8					**57**	-27	43
I_6								
	1						-20	06
	4						-38	-72
	8						**20**	-03
I_7								
	1							-06
	4							16
	8							-13

Note: Decimal points are omitted; $p < .05$ for $r > 28$; $p < .01$ for $r > 36$.

a secondary ability on which the two groups of interest differed in mean level. The results of the present study showed that such items were identified as more biased by the majority of the bias indices studied. The results were consistent across data sets varying in number of items loading on the second factor (8 or 15) and in the mean secondary ability level of the minority group ($-.5$ and -1). Items created to be biased had, on average, larger base high areas, smaller base low areas, larger signed and unsigned area indices, larger chi-square indices, and more negative difference in the *b* parameter. The only index that did not consistently identify this type of item as biased was I_7, differences in *a* parameters. This result is understandable because differences in mean level of ability are more likely to affect the difficulty parameter rather than the discriminating parameter. In fact, the ICCs of the biased items indicated that this type of bias is related to differences in difficulty level of the items in the two groups.

Although the mean item bias indices of biased items were significantly different from those of the unbiased items, not all biased items were identified as biased (using the chi-square significance test and comparison of indices to those of a baseline group). In some instances items were identified as unbiased when they were created to be biased and vice versa. An examination of these false positives and false negatives revealed that they tended to be items with more extreme thresholds (i.e., very difficult or very easy items).

Items that measure different things in groups have also been suggested as a cause of bias. In the present study, however, there was little evidence that items with different factor structures in the two groups were more biased than items with the same factor structures. The only index responsive to this type of differences across groups was I_7. This index represents differences in *a* parameters and was larger in the biased items than in a baseline comparison, being more sensitive to factor structure differences than to group ability differences.

The conclusion that items with different factor structures in two groups will not be detected as biased is a premature one since only a limited number of factor structure differences were examined. These included structures in which some items loaded on an additional factor in one group and structures in which the secondary factor for some items was different in the two groups. None of the conditions examined the effects of different loadings on the first factor, which are directly related to the *a* parameter. Also, the items included in the study did not have radically different factor structures in the two groups.

A second focus on the study was the amount of agreement among the item bias indices. In general, a great deal of agreement was found. All of the area measures were highly related. The two unsigned composite area measures (I_3 and I_5) were almost perfectly related. Linn et al. (1980) have also found similar results across the different area indices. The two parameter difference indices (I_7 and I_8) had low to moderate relationships. This result is to be expected since the two parameters are relatively independent. There was also agreement between the area indices and the parameter difference indices. As in the Shepherd et al. (1981) study, the chi-square index (I_6) consistently showed a high relationship with the unsigned area index (I_3). Difference in the *b* parameters (I_8) was highly related to all the area measures, but in particular had near perfect correlations with the signed area index (I_4).

Comparing the different indices in the unequal mean factor score conditions, the chi-square index (I_6) was always one of the best indices for identifying the items created to be biased when using the two *SD* criterion or significance test. However, the chi-square index also falsely identified other items as biased. The sum of squares index and differences in the *b* parameter were nearly as good as the chi-square index at correctly identifying biased items and also identified fewer false positives. There was a great deal of overlap in the items identified as biased with these three indices.

False positives may be reduced with the use of a two-stage procedure suggested by Marco (1977), which was not incorporated in the present study. Use of this procedure involves first estimating item parameters, calculating bias indices, identifying and deleting biased items, then estimating abilities using the remaining unbiased items. Finally, holding

ability for each individual fixed at that obtained with the unbiased items, a and b parameters are reestimated for all items and bias indices are recalculated. The rationale for this procedure is that the presence of biased items may lead to poor ability estimates, which could lead to poor identification of biased items. The present results lend support for this idea: When biased items were present, some nonbiased items were identified as biased and more of these false positives were identified in conditions in which there were more biased items. Although this "purification" procedure involves more work and computer time, it may be warranted when using those indices that tend to identify more false positives (e.g., I_6).

Although the present study provided insight into the performance of the bias indices with respect to false positives and false negatives, more research is needed to aid practitioners in choosing an index for their particular situations. The present study was limited in the type of data sets studied. Also, not all proposed bias indices were included in the study (see Hulin, Drasgow, & Parsons, 1983, and Shepherd et al., 1984, for additional indices). Questions still remain concerning the effects of including the c parameter, the effect of nonnormal distributions, and the effects of Marco's (1977) purification process.

In summary, the results of the present study supported the hypothesis of multidimensionality with lower mean ability level on secondary abilities for one group as a possible cause of item bias. On the other hand, the study did not find much support for the suggestion of items with different factor structures in two groups as a cause of bias. Regardless of the cause of bias, a great deal of agreement was found among the various item bias indices.

References

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore: Johns Hopkins University Press.

Devine, P. J., & Raju, N. S. (1982). Extent of overlap among four item bias methods. *Education and Psychological Measurement, 42,* 1049–1066.

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., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority operators. *Journal of Applied Psychology, 60,* 411–419.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. 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 for detecting test bias* (pp. 117–160). Baltimore: Johns Hopkins University Press.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge MA: Abt Books.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension* (Report No. 163). Urbana-Champaign IL: University of Illinois, Center for the Study of Reading.

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,* 159–163.

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

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

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.

Shepherd, 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.

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

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

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver, Canada: Educational Research Institute of British Columbia.

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.

Acknowledgments

Research reported in this article is part of the first author's dissertation at the University of Georgia, U.S.A. The authors thank S. Tai Chang for his assistance with the computer programming.

Author's Address

Send requests for reprints or further information to Cynthia McCauley, Center for Creative Leadership, P.O. Box P-1, Greensboro NC 27402, U.S.A.