Multidimensionality and Item Bias in Item Response Theory

T. C. Oshima, Georgia State University

M. David Miller, University of Florida

This paper demonstrates empirically how item bias indexes based on item response theory (IRT) identify bias that results from multidimensionality. When a test is multidimensional (MD) with a primary trait and a nuisance trait that affects a small portion of the test, item bias is defined as a mean difference on the nuisance trait between two groups. Results from a simulation study showed that although IRT-based bias indexes clearly distinguished multidimensionality from item bias, even with the presence of a between-group difference on the primary trait, the bias detection rate depended on the degree to which the item measured the nuisance trait, the values of MD discrimination, and the number of MD items. It was speculated that bias defined from the MD perspective was more likely to be detected when the test data met the essential unidimensionality assumption. *Index terms: item bias, multidimensionality; item response theory, item bias, mean differences, multidimensionality; multidimensionality; mean differences in IRT.*

Item bias techniques based on item response theory (IRT) rely on the property of invariance of item parameters. According to the invariance property, the same item response functions (IRFs) are obtained for a test item regardless of the trait (θ) distribution of the examinees used to estimate the item parameters (Hambleton, Swaminathan, & Rogers, 1991). In item bias research, "focal group" usually refers to the group of interest (e.g., a "minority" group), and "reference group" is the base group to which the performance of the focal group is compared (e.g., a "majority" group). Thus, when the reference and focal groups are compared, the IRFs obtained from each group would differ only as a result of sampling fluctuations, after the item parameters are adjusted onto a common scale.

This property, however, holds only when the IRT model fits the data. When the one-dimensional model does not fit the data, non-coinciding IRFs between the two groups can be expected. When IRFs differ, the probability of answering an item correctly differs between groups at a given θ level. Hence, the potential to detect item bias exists.

Model misfit due to the violation of the unidimensionality assumption has been suggested as an explanation for noncoinciding IRFs (Hunter, 1975; Miller & Linn, 1988; Oshima & Miller, 1990; Traub, 1983). Multidimensionality of a test, however, is not a sufficient condition for an item to be biased. Recently, a theory was developed to explain item bias from a multidimensional IRT perspective (Ackerman, 1991; Shealy & Stout, 1989, 1991). According to the theory, when examinees have multidimensional trait distributions that result from "primary" and "nuisance" traits, and the test items are sensitive to these differences, the differences in the conditional nuisance trait distributions between the groups of interest are the cause of potential item bias (Ackerman, 1991). In a two-dimensional perspective, one dimension represents a primary trait that is intended to be measured by the test; the second dimension represents a nuisance trait that is not intended to be measured by the test.

Following the Shealy-Stout mathematical perspective of potential bias, Ackerman listed four possible ways for conditional trait distributions to differ between two groups:

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 3, September 1992, pp. 237-248

0146-6216/92/030237-12$1.85

1. The primary trait means differ, provided there is a correlation between primary and nuisance traits.
2. The nuisance trait means differ.
3. The ratio of the variance of primary trait to that of nuisance trait is not the same for both groups.
4. The correlation of the primary and nuisance traits is not the same for both groups.

See Ackerman (1991) for mathematical explanations of the above conditions.

The second condition is of interest for this study. The mean difference on the nuisance trait is probably the most apparent source of potential bias. What is implied is that when there is no mean difference on the nuisance trait, there is no potential bias. Therefore, even when an item measures multiple traits (multidimensional items, hereafter), multidimensionality alone is not the cause of bias.

The conditions discussed above are a theoretical definition of potential item bias. The degree to which item bias detection techniques identify potentially biased items depends on other factors, such as the magnitude of the mean difference on the nuisance trait between the groups and the item direction of the item. The item direction (discussed below) determines the degree to which an item measures each of the two traits (i.e., primary and nuisance traits).

Another factor to be considered with respect to the identification of potentially biased items is the number of biased items on a test. A biased item is identified when the item measures an additional trait that is different from the trait that is recovered by unidimensional IRT calibration programs such as LOGIST and BILOG. As the number of biased items increases, the nature of the recovered trait changes. Reckase (1979), Wang (1986), and Yen (1985) have suggested both analytically and empirically that the trait obtained by unidimensional analysis applied to multidimensional data is the weighted composite of the multiple traits in the multidimensional data, with weights proportionate to the relative discriminations of the multiple traits. The weighted composite is known as the "reference composite."

Suppose that on a 40-item math test one item is a word problem, and the remaining items are computational problems. Furthermore, suppose the focal and reference groups differ in reading competency. Under these circumstances, the word problem item would show noncoinciding IRFs for the two groups, because at a given trait level (which is computational ability) differential performance would be expected favoring the group with higher reading skill. On the other hand, if 20 items were word problems, then the trait recovered by unidimensional IRT calibration would not be the same trait as in the previous example. The trait in the latter example would be the composite of computational and verbal skills. Therefore, 20 items would not show bias to the same degree as one item did in the first example.

As the number of multidimensional items increases, the assumption of unidimensionality becomes increasingly untenable. The traditional definition of unidimensionality may be too stringent, and many researchers (e.g., Harrison, 1986; Reckase, Ackerman, & Carlson, 1988; Traub, 1983) have suggested that it is unlikely that test data—especially achievement test data—will meet this assumption. Stout (1987, 1990) introduced the notion of "essential unidimensionality" in which test data can have multiple underlying traits as long as there is a dominant trait and the other traits have a relatively small influence on item scores. Program DIMTEST (Stout, Nandakumar, Junker, Chang, & Steidinger, 1991) performs hypothesis testing of essential unidimensionality. See Nandakumar (1991) for a discussion of traditional versus essential unidimensionality.

If there are biased items on a test, the test data do not meet the unidimensionality assumption in the traditional sense. If the number of biased items is small, then it is more likely that the data will meet the assumption of essential unidimensionality. However, with a larger number of biased items, even the weaker assumption of essential unidimensionality may be violated. It

is not known to what degree biased items are identified when the assumption of essential unidimensionality is violated. It was hypothesized in the present research that the detection rate of bias caused by a mean difference on the nuisance trait would decrease as the assumption of essential unidimensionality became less likely to be met.

This study had two purposes. One was to confirm empirically (using IRT-based item bias detection techniques) the theory that multidimensional items are not necessarily biased unless there is a mean difference on the nuisance trait, and also to confirm that the mean difference on the primary trait will not cause bias provided there is no correlation between the primary and the nuisance traits. The other purpose was to investigate the degree to which the theory presented above holds as a function of item direction and the number of multidimensional items on the test.

Method

Design

Data for a two-dimensional test structure were simulated in the study: θ_1 was the primary trait the test was purportedly measuring, and θ_2 was a nuisance or irrelevant trait that influenced only a small proportion of the items. The item parameters were the same for the reference and focal groups in each condition investigated. In all conditions, the θs of the reference group were normally distributed with a mean of 0 and a standard deviation (SD) of 1. Data for both the No-Bias and Bias conditions were simulated using mean differences on θ_2. For the No-Bias condition, the focal and reference groups had the same distribution of scores on θ_2. For the Bias condition, the focal and reference groups had different means on θ_2. The mean difference was $\bar{\theta}_{2A} - \bar{\theta}_{2B} = .5$, where $\bar{\theta}_{2A}$ and $\bar{\theta}_{2B}$ are the means on θ_2 for Group A (focal group) and Group B (reference group), respectively. Linn & Drasgow (1987) reported that mean θ differences between Black and White test takers were typically 1 SD. Because smaller differences might be expected for other subpopulations (e.g., gender differences), a more conservative difference of .5 SD was used in this study. This difference might be expected on the primary trait (e.g., math ability) or the nuisance trait (e.g., reading ability on word problems). The mean difference of .5 on θ_2 was defined here as biased, because the test was constructed to measure θ_1 only.

For both the No-Bias and Bias conditions, two factors were considered. One factor, the between-group mean difference on θ_1, had two levels: $\bar{\theta}_{1A} - \bar{\theta}_{1B} = 0.0$ and $\bar{\theta}_{1A} - \bar{\theta}_{1B} = .5$. The second factor—the percentage of items that were multidimensional (MD) with respect to θ_2 had three levels: 5%, 10%, and 20%. Thus, 5%, 10%, or 20% of all the items (the last 2, 4, and 8 items of the 40-item test, respectively) measured two underlying dimensions or traits. These percentages coincide with those obtained for published tests when item bias studies have been conducted. For example, Drasgow (1987) found, in his investigation of item bias in ACT assessment tests in various subjects, that 5% to 29% of the items were biased based on Lord's χ^2 technique. The correlation between θ_1 and θ_2 was assumed to be 0 in all conditions.

The three factors were crossed and resulted in 12 ($2 \times 2 \times 3$) different conditions: two levels of $\bar{\theta}_{2A} - \bar{\theta}_{2B}$ (No-Bias vs. Bias), two levels of $\bar{\theta}_{1A} - \bar{\theta}_{1B}$ (between-group difference), and the three levels of the number of MD items. The 12 conditions were replicated 10 times. An additional factor of interest, the item direction (see below), was embedded throughout the condition; in each condition, MD items had different levels of item direction ranging from measuring mostly θ_1 to measuring more θ_2 than θ_1.

Data Generation

θ distributions. Each simulated dataset included two groups of simulated examinees: 1,000 examinees from a reference group and 1,000 examinees from a focal group. The θ values were randomly generated from a normal [N(0,1)] distribution using the RANNOR function in SAS (SAS Institute Inc., 1990). The mean of the

distribution was altered by adding .5 to the generated θs when it was applicable.

Model and item parameters. The MD data were generated using a compensatory MD two-parameter logistic (M2PL) model (Reckase & McKinley, 1991). The M2PL model is

$$P(x_{ij} = 1 \mid \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{\exp(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}{1 + \exp(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)} , \quad (1)$$

where x_{ij} is the 0 or 1 score on item i for person j,
$\mathbf{a}_i$ is the vector of item discrimination parameters,
d_i is a scalar parameter related to the difficulty of the item, and
$\boldsymbol{\theta}_j$ is the vector of trait parameters for person j.

Reckase (1985) defined an MD item difficulty (MID) parameter by

$$\text{MID}_i = -d_i \Big/ \left[\sum_{k=1}^{m} (a_{ik})^2\right]^{.5} , \quad (2)$$

where a_{ik} is the kth element of $\mathbf{a}_i$. Reckase & McKinley (1991) also defined an MD discrimination parameter (MDISC) as

$$\text{MDISC}_i = \left[\sum_{k=1}^{m} (a_{ik})^2\right]^{.5} . \quad (3)$$

Item parameters for this study were selected to reflect actual test data. MID parameters were selected from a normal distribution with mean = 0 and SD = 1 within a range of −2 to 2, and MDISC parameters were selected randomly from a lognormal distribution with mean 1.13 and SD .60.

Item direction. In the M2PL model, item directions (α_{i1}) determine the weighted composite of traits measured by an item. The angles can be determined using the direction cosines given by

$$\cos\alpha_{ik} = a_{ik} \Big/ \left[\sum_{k=1}^{m} (a_{ik})^2\right]^{.5} , \quad (4)$$

where a_{ik} is the kth element of $\mathbf{a}_i$. In a two-dimensional space, if an item measures only θ_1, then α_{i1} is 0°; if an item measures only θ_2, α_{i1} is 90°. α_{i1} can be any value from 0° to 90° depending on the degree to which an item measures the two traits. If α_{i1} = 45°, for example, the item measures θ_1 and θ_2 equally. In the present study, all but 5% (10% or 20%) of the items had α_{i1} = 0°. This is the situation in which all the items measure θ_1 and only a small percentage of items measure both θ_1 and θ_2.

When 10% of the items (i.e., four items) were MD, α_{i1} was 15° for the first of the four items, 30° for the second, 45° for the third, and 60° for the fourth. Without loss of generality, these were the last four items on the test. When 20% of the items were MD, the item parameters for Items 33, 34, 35, and 36 were equal to the item parameters for Items 37, 38, 39, and 40, respectively (which were the parameters for the MD items in the test in which 10% of the items were MD). When 5% of the items were MD, the item parameters for Items 39 and 40 were the same as those items were in the test in which 10% of the items were MD. Items 37 and 38 had item parameters equal to those for Items 1 and 2. Note that in evaluating the effect of the number of MD items, it is reasonable to directly compare the 10% to the 20% condition; however, comparison of the 5% to the 10% condition should be made with caution, because the MD items in the 5% condition had higher average α_{i1} than the 10% or 20% condition.

The descriptive statistics for item parameters with 10% of the items MD were:

MDISC ranged from .53 to 3.12, with mean = 1.18 and SD = .58;
a_{i1} ranged from .47 to 3.12, with mean = 1.15 and SD = .58;
a_{i2} for Items 37–40 ranged from .19 to 1.24, with mean = .70 and SD = .45;
MID ranged from −1.58 to 1.84 with mean = .17 and SD = .79; and
d_i ranged from −3.42 to 2.82, with mean = −.25 and SD = 1.08.

Item data. The probability of answering an item correctly was calculated using Equation 1, and the result was compared to random numbers

generated from a uniform distribution. If the random number was less than or equal to the probability, $x_{ij} = 1$; otherwise, $x_{ij} = 0$.

Analysis

The responses of 1,000 simulees from each subgroup to the 40 items in each condition were analyzed using a unidimensional (UD) two-parameter logistic model by PC-BILOG (Mislevy & Bock, 1986) with default priors—a log-normal prior on the discrimination estimates and no prior on the difficulty estimates. The item parameter estimates then were placed on the same scale using the means and the SDs of the item difficulty estimates from both groups (see Marco, 1977), and four item bias indexes were computed: signed area (SA), unsigned area (UA), signed sum of squares (SSOS), and unsigned sum of squares (USOS). The four indexes are described in detail in Shepard, Camilli, & Williams (1985). Because the distributions of these indexes are unknown, an item is identified as biased if the index value exceeds the baseline mean by two SDs. In practice, the baseline is created by taking two samples from the same population (normally from the reference group). Previous research (Oshima, 1989) has shown, however, that the baseline method is not stable; the criterion can fluctuate depending on the sample. Thus, in the present study, the baseline was replicated five times. Then, the criterion value obtained from each replication was averaged across the five replications. The criterion values used for the present study are reported in Table 1.

Table 1
Criterion Values for Determining Bias for Each Index and the Percentage of Biased Items Conditions

Index and Percentage of MD Items	Criterion
SA	
5%	±.173
10%	±.205
20%	±.201
UA	
5%	.262
10%	.277
20%	.293
SSOS	
5%	±1.13
10%	±1.23
20%	±1.55
USOS	
5%	1.56
10%	1.60
20%	2.03

Results

Table 2 summarizes the results of the analyses of the 10 replications of the 12 conditions for SA, UA, SSOS, and USOS. For each index, the mean number of items that exceeded the criterion, the mean proportion of items that exceeded the criterion, and the mean of the mean index value over 10 replications are reported for UD and MD items.

No-Bias Condition

Table 2 shows that MD items were not necessarily identified as biased when θ_2 had the same distribution for the reference and focal groups ($\bar{\theta}_{2A} - \bar{\theta}_{2B} = 0$). For USOS, for example, the proportions of items that exceeded the criterion for the MD items were 0, .05, and .01 for the 5%, 10%, and 20% conditions, respectively. Those values for the UD items were .08, .05, and .03, for the 5%, 10%, and 20% conditions, respectively. The results show that false positives (i.e., unbiased items with index values that exceeded the criterion) occurred no more frequently in the MD items than in the UD items. The same trend was observed even when there was a between-group difference on θ_1.

Bias Condition

In contrast to the No-Bias condition, when the reference group and the focal group had different means on θ_2 ($\bar{\theta}_{2A} - \bar{\theta}_{2B} = .5$), those MD items with bias were identified as biased at much higher rates. For all the indexes (SA, UA, SSOS, USOS),

Table 2
Mean Number of Items That Exceeded the Criterion (n), Mean Proportion of Items That Exceeded the Criterion (P), and Mean Value of Item Bias Index (M) for UD and MD Items, for SA, UA, SSOS, and USOS Indexes, and for No-Bias ($\bar{\theta}_{2A} - \bar{\theta}_{2B} = 0$) and Bias ($\bar{\theta}_{2A} - \bar{\theta}_{2B} = .5$) Conditions

Percentage of MD Items, UD or MD, and Number of Items	No Between-Group Difference on θ_1 ($\bar{\theta}_{1A} - \bar{\theta}_{1B} = 0$)						Between-Group Difference on θ_1 ($\bar{\theta}_{1A} - \bar{\theta}_{1B} = .5$)					
	No Bias			Bias			No Bias			Bias		
	n	P	M	n	P	M	n	P	M	n	P	M
SA												
5%: UD, 38												
M	3.5	.09	.005	3.8	.10	.025	2.8	.07	.003	4.2	.11	.028
SD	1.4	.04	.009	1.4	.04	.009	1.4	.04	.008	1.6	.04	.014
5%: MD, 2												
M	.1	.05	.032	2.0	1.00	−.371	.1	.05	−.006	1.9	.95	−.348
SD	.3	.16	.052	0.0	0.00	.040	.3	.16	.042	.3	.16	.008
10%: UD, 36												
M	.9	.02	.003	1.6	.04	.033	1.1	.03	.001	2.0	.06	.045
SD	1.0	.03	.008	1.4	.04	.015	1.0	.03	.012	1.3	.04	.017
10%: MD, 4												
M	.2	.05	−.006	2.4	.60	−.254	.1	.02	.003	2.5	.62	−.268
SD	.4	.11	.042	.5	.13	.048	.3	.08	.044	.5	.13	.048
20%: UD, 32												
M	1.4	.04	.002	2.5	.08	.052	1.6	.05	−.001	2.6	.08	.061
SD	.5	.02	.008	1.4	.04	.008	1.0	.03	.020	1.7	.05	.013
20%: MD, 8												
M	.3	.04	.010	4.3	.54	−.213	.9	.11	.027	4.0	.50	−.188
SD	.7	.08	.024	.7	.08	.030	.6	.07	.039	.7	.08	.033
UA												
5%: UD, 38												
M	2.4	.06	.130	3.0	.08	.136	1.6	.04	.126	2.8	.07	.133
SD	1.5	.04	.011	1.7	.04	.014	1.1	.03	.005	1.8	.05	.014
5%: MD, 2												
M	0.0	0.00	.131	2.0	1.00	.374	0.0	0.00	.135	1.6	.80	.360
SD	0.0	0.00	.041	0.0	0.00	.046	0.0	0.00	.041	.5	.26	.067
10%: UD, 36												
M	1.3	.04	.126	2.6	.07	.145	1.7	.05	.130	2.5	.07	.145
SD	.7	.02	.005	1.7	.05	.025	1.3	.04	.015	1.9	.05	.023
10%: MD, 4												
M	.1	.02	.135	1.9	.47	.297	.2	.05	.132	2.1	.52	.317
SD	.3	.08	.041	.6	.14	.042	.4	.11	.037	.6	.14	.036
20%: UD, 32												
M	.7	.02	.120	2.9	.09	.147	1.0	.03	.126	1.7	.05	.144
SD	.8	.03	.006	2.1	.07	.025	1.1	.03	.015	1.3	.04	.021
20%: MD, 8												
M	.2	.02	.130	3.4	.42	.263	.4	.05	.158	2.5	.31	.236
SD	.4	.05	.027	1.2	.15	.033	.7	.09	.029	1.3	.16	.026
SSOS												
5%: UD, 38												
M	3.5	.09	.02	3.1	.08	.15	2.7	.07	.02	3.3	.09	.12
SD	1.5	.04	.09	1.4	.04	.06	1.1	.03	.06	1.3	.03	.06

continued on the next page

Table 2, continued

Mean Number of Items That Exceeded the Criterion (n), Mean Proportion of Items That Exceeded the Criterion (P), and Mean Value of Item Bias Index (M) for UD and MD Items, for SA, UA, SSOS, and USOS Indexes, and for No-Bias ($\bar{\theta}_{2A} - \bar{\theta}_{2B} = 0$) and Bias ($\bar{\theta}_{2A} - \bar{\theta}_{2B} = .5$) Conditions

Percentage of MD Items, UD or MD, and Number of Items	No Between-Group Difference on θ_1 ($\bar{\theta}_{1A} - \bar{\theta}_{1B} = 0$)						Between-Group Difference on θ_1 ($\bar{\theta}_{1A} - \bar{\theta}_{1B} = .5$)					
	No Bias			Bias			No Bias			Bias		
	n	P	M	n	P	M	n	P	M	n	P	M
5%: MD, 2												
M	0.0	0.00	.08	2.0	1.00	−2.88	0.0	0.00	−.06	1.6	.80	−2.89
SD	0.0	0.00	.29	0.0	0.00	.78	0.0	0.00	.21	.5	.26	.81
10%: UD, 36												
M	1.6	.04	.02	4.0	.11	.23	2.1	.06	.02	2.9	.08	.24
SD	1.3	.04	.06	2.3	.06	.14	1.7	.05	.06	2.1	.06	.13
10%: MD, 4												
M	.2	.05	−.06	2.2	.55	−2.18	.2	.05	−.05	2.5	.62	−2.40
SD	.4	.11	.21	.6	.16	.88	.4	.11	.21	.5	.13	.58
20%: UD, 32												
M	1.4	.04	−.03	2.5	.08	.34	1.1	.03	−.04	2.5	.08	.33
SD	.7	.02	.05	1.7	.05	.14	1.0	.03	.08	1.9	.06	.10
20%: MD, 8												
M	.2	.02	.01	3.3	.41	−1.56	.2	.02	.15	3.0	.37	−1.33
SD	.4	.05	.16	1.2	.14	.34	.4	.05	.19	1.1	.13	.37
USOS												
5%: UD, 38												
M	3.0	.08	.56	3.6	.09	.60	2.2	.06	.51	3.0	.08	.58
SD	1.3	.04	.10	2.0	.05	.13	1.1	.03	.05	1.9	.05	.09
5%: MD, 2												
M	0.0	0.00	.45	2.0	1.00	2.89	0.0	0.00	.53	1.6	.80	2.94
SD	0.0	0.00	.23	0.0	0.00	.79	0.0	0.00	.22	.5	.26	.86
10%: UD, 36												
M	1.7	.05	.51	3.8	.11	.69	2.2	.06	.55	3.4	.09	.67
SD	1.2	.03	.05	2.1	.06	.23	1.4	.04	.14	2.3	.06	.20
10%: MD, 4												
M	.2	.05	.53	1.9	.47	2.36	.2	.05	.51	2.1	.52	2.59
SD	.4	.11	.22	.7	.18	.86	.4	.11	.23	.6	.14	.62
20%: UD, 32												
M	1.0	.03	.48	2.9	.09	.75	1.1	.03	.52	1.8	.06	.71
SD	1.1	.03	.04	2.4	.07	.25	1.2	.04	.13	1.6	.05	.21
20%: MD, 8												
M	.1	.01	.50	3.0	.37	1.82	.4	.05	.69	2.3	.29	1.47
SD	.3	.04	.20	1.2	.14	.43	.7	.09	.22	1.6	.20	.35

100% of the items were identified as biased when there was no between-group difference with 5% MD items. This higher detection rate also is explained by the higher mean index values for the MD and biased items for the 5% condition. For example, the mean USOS value for the 38 UD items was .60, and that of the two MD items was 2.89. The same trend was observed even when there was a between-group difference on θ_1. The detection rates decreased slightly; for example, they decreased from 100% to 80% for USOS in the 5% condition. However, the mean index values were clearly inflated when the items were MD. For example, the mean USOS value for the 38 UD items was .58, and that of the two MD items was 2.94. These values were comparable to

those obtained in the no-between-group-difference condition. The false positive rate (i.e., the detection rate for unbiased items) increased slightly compared with the No-Bias condition. However, the detection rate for the UD items was substantially lower than that for MD items with a mean difference on θ_2.

The proportion of biased items detected correctly decreased as the number of biased items increased. For example, across all the indexes the detection rate ranged from .47 to .62 with a mean of .55 when 10% of the items were biased, and from .29 to .54 with a mean of .40 when 20% of the items were biased. For the 5% condition, the detection rate ranged from .80 to 1.00.

An additional analysis was conducted to counterbalance the effect of α_{i1} in the 5% condition. The Bias condition with no between-group difference was replicated 10 times with $\alpha_{i1} = 15°$ and 30° for Item 39 and Item 40, respectively. Note that in the original 10 replications of the 5% condition, $\alpha_{i1} = 45°$ and 60° for Items 39 and 40, respectively. As expected with smaller α_{i1}s, the detection rate decreased. The detection rates for SA, UA, SSOS, and USOS were .55, .35, .45, and .35, respectively. The mean detection rate across 20 replications (i.e., 10 replications with $\alpha_{i1} = 45°$ and 60° and 10 replications with $\alpha_{i1} = 15°$ and 30°) for the 5% condition was .67. These results suggest that, after controlling for the α_{i1} effect, there is a general trend: As the number of biased items increases, the detection rate decreases.

Unidimensionality

To examine the essential unidimensionality of the data, 10 datasets for the reference group from each condition (5%, 10%, and 20%) were tested for essential unidimensionality using DIMTEST (Stout et al., 1991). Of 10 replications for the data with 20% MD, five datasets were rejected when the null hypothesis of essential unidimensionality was tested at $\alpha = .005$ (.05/10). Six datasets were rejected at $\alpha = .05$. For the 10% condition, none of the datasets was rejected at $\alpha = .005$, but two were rejected at $\alpha = .05$. Finally, for the 5% condition, none of the datasets was rejected at $\alpha = .005$, but one was rejected at $\alpha = .05$. These results imply that the higher the number of MD items, the greater the likelihood that essential unidimensionality will be violated. Furthermore, these results suggest that bias introduced by mean differences on θ_2 is more likely to be detected when the essential unidimensionality assumption is met.

Effect of Item Directions

The effect of α_{i1} was examined by counting how many times each item was identified as biased in the 10 replications. The results from the Bias condition with no between-group difference for the USOS index are reported in Table 3. Similar trends were observed for the condition with the between-group difference, and for the other three

Table 3
Item Characteristics and Number of Times Each Item Was Identified as Biased in the 10 Replications as a Function of α_{i1} for the 5%, 10%, and 20% Bias Conditions

				Number of Identifications		
Item	α_{i1}	MDISC	MID	5%	10%	20%
33	15	.73	.12			0
34	30	1.07	1.38			1
35	45	1.75	.14			6
36	60	.94	1.28			6
37	15	.73	.12		0	1
38	30	1.07	1.38		3	0
39	45	1.75	.14	10	9	8
40	60	.94	1.28	10	5	6

indexes. As expected, the general trend is that the larger the α_{i1}, the better the detection. However, the trend was not consistent. For example, Item 39 had a higher detection rate overall than Item 40, although Item 40 had a larger α_{i1} than Item 39.

Effect of MDISC and MID

Because of the results for item directions, other characteristics of the item—MDISC and MID—were investigated. By observing Table 3, it was hypothesized that the higher MDISC would lead to a higher detection rate. To confirm this hypothesis, new datasets were generated for the 5%, 10%, and 20% conditions. The same procedures were followed as before, except that the MD items had two levels of MDISC, and MID and α_{i1} were held constant. No replications were done for this additional analysis. Two extreme values of MDISC were selected: One value was highly discriminating (MDISC = 2.0), and the other was less discriminating (MDISC = .5). MID was set to 0, and α_{i1} was 45° for all the MD items.

The USOS values for each item are shown in Table 4. The effect of MDISC was evident. Biased items with higher MDISC were more likely to be identified as biased items. The mean USOS values with MDISC of .5 were .43, .62, and .85 for the 5%, 10%, and 20% conditions, respectively. On the other hand, the mean USOS values with MDISC of 2.0 were 3.98, 3.71, and 2.07 for the 5%, 10%, and 20% conditions, respectively.

The effect of MID also was investigated with other datasets. No consistent trend was observed for the effect of MID when other item characteristics were held constant.

The behavior of the four indexes was similar in all the conditions. For signed indexes, the direction of bias was explicitly indicated by the negative mean index values when the bias was embedded in the MD items.

Discussion

The results demonstrated that bias caused by a mean difference on the nuisance trait between two groups was identified by IRT-based bias indexes, provided that the number of biased items was small and the item direction was fairly large. These results suggest that the IRT-based indexes are a powerful way of detecting bias, because they can successfully detect multidimensional items with bias but do not detect multidimensional items without bias. Furthermore, they do not confound between-group differences on the primary trait. This is a useful property, because it is likely that in a bias study between-group differences will be observed on the primary trait the test measures.

The extent of the properties of IRT-based indexes depends on the number of biased items, the item direction, and multidimensional item discrimination. When a larger proportion of the test is biased, the power of item bias detection techniques decreases. Thus, item bias detection

Table 4
USOS Values as a Function of MDISC
for the 5%, 10%, and 20% Bias Conditions

				USOS		
Item	α_{i1}	MDISC	MID	5%	10%	20%
33	45	.5	0			1.48
34	45	2.0	0			3.77*
35	45	.5	0			1.03
36	45	2.0	0			.78
37	45	.5	0		.73	.17
38	45	2.0	0		3.47*	1.85*
39	45	.5	0	.43	.50	.72
40	45	2.0	0	3.98*	3.95*	1.87*

*Item identified as potentially biased.

will be effective when idiosyncrasies of an item cause multidimensionality and when there is a mean difference on the second trait. On the other hand, item bias detection will be less effective when mean differences on a second trait have a more pervasive effect on the test, as might be expected with bilingual subpopulations or subpopulations with instructional differences (Miller & Linn, 1988). Another simulation study (Oshima, 1989) showed that when all the items are multidimensional with two dominant traits, the mean difference on the nuisance trait did not cause an excessive number of items to be identified as biased.

Item bias due to a mean difference on the nuisance trait between two groups and violation of the assumption of unidimensionality are closely related. When a small number of items is multidimensional, the test is essentially unidimensional. However, as the number of multidimensional items increases, the assumption of essential unidimensionality becomes less tenable. In this study, the tests with 5% multidimensional items were shown to be essentially unidimensional; however, the hypothesis of essential unidimensionality was rejected half of the time for the tests with 20% multidimensional items. Therefore, test data should be tested for dimensionality before applying item bias analyses. If the assumption of essential dimensionality is rejected, then IRT-based item detection indexes that are based on a unidimensional IRT model may not be appropriate.

Several researchers have suggested iterative IRT item bias detection techniques when a fairly large number of items are expected to be biased (Miller & Oshima, in press; Park & Lautenschlager, 1990). In the iterative approach, the test data are "purified" by removing biased items. In essence, by removing the biased items, the remaining test approaches a more unidimensional structure. The more direct approach to obtaining the unidimensional structure was suggested by Shealy & Stout (1989) and Shealy, Stout, & Rossos (1991). In this approach, a "valid" subtest, which is as unidimensional as possible, can be selected by a practitioner, and items suspected to be biased can be tested against the valid subtest. The selection of the valid subtest can be rather subjective. More research is needed in this area, however.

As expected, in this study the biased items were identified as biased more often as the degree to which the item measured the nuisance trait increased. Multidimensional discrimination (MDISC) also was found to be a factor that influenced the detection rate of biased items. When the value of MDISC increased, the discrimination power for both traits (i.e., a_{i1} and a_{i2}) increased. As a_{i2} (i.e., the discrimination power on the nuisance trait) increased, the item with higher a_{i2} became more sensitive to a distributional difference on the nuisance trait, thus resulting in the higher detection rate of biased items.

The generalizability of the results of this study is limited by the range of values for item parameters, the trait distributions for each group, and the dimensionality structure. In the present study, the mean differences between groups on the primary and nuisance traits were set at .5. In practice, there may be a difference as large as 1.0 between the focal and reference groups on the primary trait (Linn & Drasgow, 1987). As described above, there are other types of bias in addition to mean differences between the nuisance traits. In this study, the dimensionality structure was limited to two dimensions. There was only one nuisance trait. However, it is also plausible to model bias from multiple nuisance traits. Further research is needed to delineate the relationship between multidimensionality and item bias under various configurations of multidimensionality with various types of item bias.

References

Ackerman, T. A. (1991, April). *A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Hambleton, R. K., Swaminathan, H., & Rogers, J. H.

(1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violation of the unidimensionality assumption. *Journal of Educational Measurement, 11,* 91–115.

Hunter, J. E. (1975, December). *A critical analysis of the use of item means and item test correlations to determine the presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education Conference on Test Bias, Annapolis MD.

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice, 6(2),* 13–17.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 15,* 139–160.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25,* 205–219.

Miller, M. D., & Oshima, T. C. (in press). Two-stage estimation of item bias. *Applied Psychological Measurement.*

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

Nandakumar, R. (1991). Traditional versus essential dimensionality. *Journal of Educational Measurement, 28,* 99–117.

Oshima, T. C. (1989). *The effect of multidimensionality on item bias detection based on item response theory.* Unpublished doctoral dissertation, University of Florida, Gainesville.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. *Journal of Educational Measurement, 27,* 273–283.

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.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193–203.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373.

SAS Institute Inc. (1990). *The SAS system for personal computers: Release 6.04* [Computer program]. Cary NC: Author.

Shealy, R., & Stout, W. (1989, April). *A procedure to detect test bias presented simultaneously in several items.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Shealy, R., & Stout, W. (1991). *An item response theory model for test bias* (Tech. Rep. No. 4421-548). Champaign IL: University of Illinois, Department of Statistics.

Shealy, R., Stout, W., & Rossos, L. (1991). *SIBTEST manual* [Computer program manual]. Champaign IL: Department of Statistics, University of Illinois.

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

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Stout, W., Nandakumar, R., Junker, B., Chang, H., & Steidinger, D. (1991). *DIMTEST* [Computer program]. Champaign IL: Department of Statistics, University of Illinois.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). Vancouver: Educational Research Institute of British Columbia.

Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at the ONR contractor's conference, Gatlinburg TN.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50,* 399–410.

Acknowledgments

The authors thank the editor and two anonymous reviewers for their helpful comments, James Algina and Linda Crocker for their useful insights at the early stages of this study, and Claudia Flowers for her assistance in data analyses.

Author's Address

Send requests for reprints or further information to T. C. Oshima, Dept. of Educational Foundations, Georgia State University, University Plaza, Atlanta GA 30303, U.S.A.