The Influence of Conditioning Scores In Performing DIF Analyses

Terry A. Ackerman and John A. Evans
University of Illinois

The effect of the conditioning score on the results of differential item functioning (DIF) analyses was examined. Most DIF detection procedures match examinees from two groups of interest according to the examinees' test score (e.g., number correct) and then summarize the performance differences across trait levels. DIF has the potential to occur whenever the conditioning criterion cannot account for the multidimensional interaction between items and examinees. Response data were generated from a two-dimensional item response theory model for a 30-item test in which items were measuring uniformly spaced composites of two latent trait parameters, θ_1 and θ_2. Two different DIF detection methods—the Mantel-Haenszel and simultaneous item bias (SIBTEST) detection procedure—were used for three different sample size conditions. When the DIF procedures were conditioned on the number-correct score or on a transformation of θ_1 or θ_2, differential group performance followed hypothesized patterns. When the conditioning criterion was a function of both θ_1 and θ_2 (i.e., when the complete latent space was identified), DIF, as theory would suggest, was eliminated for all items. *Index terms: construct validity, differential item functioning, item bias, Mantel-Haenszel procedure, SIBTEST.*

The purpose of most standardized achievement tests is to distinguish between levels of proficiency. To properly achieve a rank ordering, all the items that contribute to the total score must discriminate between levels of the same trait or the same composite of multiple traits. That is, ordering is a unidimensional concept.

When items in a test are capable of distinguishing between levels of different traits, differential item functioning (DIF) will occur when groups of examinees differ in their distribution on these multiple traits and this multidimensional interaction between examinees and items is summarized as a single score. The main source of DIF is that the ordering/matching/conditioning criterion does not account for the distributional differences between groups in the complete latent trait space. The viewpoint that DIF is the result of using a single score when the latent trait space is multidimensional is more formally developed in Shealy & Stout's (1993a, 1993b) multidimensional item response theory (IRT) model of test bias.

For example, consider a testing situation in which two examinees, A and B, have the two-dimensional θ_1, θ_2 trait levels $(2, -2)$ and $(-2, 2)$, respectively. Assume that for a given test, the observed score represents a composite in which θ_1 and θ_2 are equally weighted. In this case, both examinees would receive the same trait estimate, even though their true trait levels are exactly opposite. Consider further two items in which Item 1 measures only θ_1, and Item 2 measures only θ_2. Even though both examinees have the same unidimensional trait level estimate, Examinee A would have a much greater probability of answering Item 1 correctly than Examinee B. Likewise, Examinee B would have a much greater probability of answering Item 2 correctly than Examinee A. There is no single composite score that would equalize the probability of correct response for both of these hypothetical individuals. Only if multiple scores were reported would the true trait differences be captured.

DIF occurs if examinee groups have different trait distributions for traits that are not intended to be measured and the test results are summarized as a single score that does not accurately represent the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 4, December 1994, pp. 329–342

0146-6216/94/040329-14$1.95

examinees' profiles of traits (i.e., the complete latent trait space). To account for test multidimensionality, some researchers (Shin, 1992; Zwick & Ercikan, 1989) have attempted to better account for the latent trait space by conditioning on multiple trait scores that were logically related to the underlying response process of the test items. In both studies (Shin; Zwick & Ercikan) the additional conditioning did not prove successful.

Shealy & Stout (1993a) presented a version of the DIF detection computer program SIBTEST in which an item was labeled as a DIF item only if two conditioning scores indicated DIF. This procedure has been demonstrated to have a Type I error rate close to the nominal level; it also has been shown to be capable of identifying generated DIF items (Ackerman & Evans, 1992).

The purpose of this research study was to demonstrate that the results of DIF analyses that rely on a conditioning score can be quite different depending on the conditioning variable that is selected. Unlike previous real data studies, this study used generated multidimensional data; thus, all of the true traits that produced the observed responses were known.

DIF Detection Methodology

Although there has been a proliferation of methods to detect DIF (Millsap & Everson, 1993), this study used the Mantel-Haenzsel (MH) procedure (Holland & Thayer, 1988) and Shealy & Stout's (1993b) simultaneous item bias (SIBTEST; henceforth referred to as SIB) procedure. Both of these procedures are nonparametric, require no model calibration, and have been shown to be effective (Ackerman & Evans, 1992). However, they can be studied from an IRT framework, and as such can be explained within a multidimensional IRT context.

The MH Procedure

To compute MH for an item *i*, two groups of interest—usually denoted as the reference and focal groups—are identified; the focal (F) group is typically a minority group, and the reference (R) group is frequently a nonminority group. Examinees from each group are matched according to their number-correct (NC) score. A 2 × 2 contingency table is created (for each possible score category) in which the frequency of correct and incorrect answers for each group are recorded along with the marginal and total frequencies. The table for the *j*th score category has the form shown in Table 1, where an item score of 1 indicates a correct response and 0 indicates an incorrect response. Summing over the contingency tables for an item *i* and using a continuity correction, the MH statistic is given by

$$\mathrm{MH}_i = \frac{\left[\left|\sum_j A_j - \sum_j \mathrm{E}(A_j)\right| - \frac{1}{2}\right]^2}{\sum_j \mathrm{Var}(A_j)} , \tag{1}$$

where the expected value of the cell A frequency is

$$\mathrm{E}(A_j) = \frac{N_{\mathrm{R}j} N_{1.j}}{N_{..j}} , \tag{2}$$

and the variance of the cell A frequencies can be computed as

$$\mathrm{Var}(A_j) = \frac{N_{\mathrm{R}j} N_{\mathrm{F}j} N_{1.j} N_{0.j}}{(N_{..j})^2 (N_{..j} - 1)} . \tag{3}$$

To remove the artificial effect of item impact (i.e., when the focal and reference group examinees differ in their distributions of the intended-to-be-measured trait), the suspect item score must be included as part of the conditioning score.

Table 1
2 × 2 Contingency Table at the *j*th Score Category

	Item Score		
Group	1	0	Total
Reference (R)	A_j	B_j	$N_{\mathrm{R}j}$
Focal (F)	C_j	D_j	$N_{\mathrm{F}j}$
Total	$N_{1.j}$	$N_{0.j}$	$N_{..j}$

The MH statistic can be used to test the null hypothesis that for each score category *j* the odds of a reference group examinee answering the item correctly equals the odds that a focal group examinee will answer the item correctly (Holland &

Thayer, 1988). Specifically, if p_{Rj} and p_{Fj} are the probabilities of a reference and focal group examinee answering the item correctly, respectively, and q_{Rj} and q_{Fj} are the probabilities of a reference and focal group examinee answering the item incorrectly, respectively,

$$H_0 : \frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}}, \quad j = 1, \ldots, K , \tag{4}$$

is tested against the alternative of uniform DIF,

$$H_A : \frac{p_{Rj}}{q_{Rj}} = \alpha \frac{p_{Fj}}{q_{Fj}}, \quad \alpha \neq 1, \quad j = 1, \ldots, K , \tag{5}$$

where H_0 is the null hypothesis, H_A is the alternative hypothesis, and α is the common odds ratio in the K 2×2 tables. Uniform DIF occurs when the rescaled, unidimensional item response functions differ only in difficulty. When H_0 is true, MH is distributed approximately as χ^2 with 1 degree of freedom.

The SIB Procedure

The SIB statistic, B_U, is computed in a manner somewhat similar to what is called the standardization procedure (Dorans & Kulick, 1986) but with several important differences (see Shealy & Stout, 1993b, for a more detailed description of B_U). First, its computation requires that the practitioner divide the N-item test under consideration into an n-item valid subtest (possibly containing all of the items except the one item suspected of DIF) and a set of $N - n$ suspect item(s). For example, the valid items can be identified as the items that weight most highly on a particular factor of a factor analysis of the item tetrachoric matrix, or are identified using a hierarchical cluster analysis (Roussos, Stout, & Marden, 1993) or by cognitive considerations. [Note that Shealy & Stout's (1993a) valid subtest is somewhat comparable to the collection of nonDIF items identified using a purification approach with the MH procedure (Dorans & Holland, 1993).] The remaining items are classified as suspect items and can be tested one at a time or collectively.

Once the test is split into these two categories, the total score on the suspect item(s),

$$Y = \sum_{i=n+1}^{N} U_i , \tag{6}$$

and the valid subtest score,

$$X = \sum_{i=1}^{n} U_i , \tag{7}$$

are computed from the dichotomous item scores, U_i. The statistics $\bar{Y}_{Rh}$ and $\bar{Y}_{Fh}$, representing the average Y for all examinees attaining a valid subtest score $X = h$ ($h = 0, 1, 2, \ldots, n$), are calculated for the reference and focal groups, respectively. To remove the source of impact, a simple true-score-theory-based regression correction is used to obtain adjusted values $\bar{Y}^*_{Rh}$ and $\bar{Y}^*_{Fh}$. Shealy & Stout (1993a) defined a model-based parameter measuring the amount of unidirectional (noncrossing) DIF present. An estimate $\hat{\beta}_U$ of B_U is defined as

$$\hat{\beta}_U = \sum_{h=0}^{n} \hat{p}_h \left(\bar{Y}^*_{Rh} - \bar{Y}^*_{Fh} \right) , \tag{8}$$

where

$$\hat{p}_h = \frac{\left(G_{Rh} + G_{Fh} \right)}{\sum_{j=0}^{n} \left(G_{Rj} + G_{Fj} \right)} , \tag{9}$$

where G_{Rh} and G_{Fh} are the number of examinees in the reference and focal groups, respectively, with the same valid score $X = h$.

The test statistic is given by

$$B_U = \frac{\hat{\beta}_U}{\hat{\sigma}\left(\hat{\beta}_U \right)} . \tag{10}$$

The denominator of this expression is the estimated standard error of $\hat{\beta}_U$ and is computed as

$$\hat{\sigma}\left(\hat{\beta}_U \right) = \left\{ \sum_{h=0}^{n} \hat{p}_k^2 \left[\frac{1}{G_{Rh}} \hat{\sigma}^2 (Y|h, \mathrm{R}) + \frac{1}{G_{Fh}} \hat{\sigma}^2 (Y|h, \mathrm{F}) \right] \right\}^{1/2} , \tag{11}$$

where the σ^2s are the empirical variances for cell h for the suspect test scores, $h = 0, \ldots, n$.

The test statistic has an approximate N(0,1) distribution when no DIF is present (i.e., $\beta_U = 0$). Thus, the hypothesis of testing DIF against the focal group can be stated as H_0: $\beta_U = 0$ versus H_A: $\beta_U > 0$.

Method

The Two-Dimensional IRT Model

The DIF simulation used a two-dimensional IRT model. This approach is more valid and realistic than using a unidimensional IRT model and assigning different generating item parameters to each group depending on the direction and amount of DIF. Ackerman (1992) outlined ways in which the underlying trait distributions could produce differences in rescaled unidimensional item response functions for the two groups of interest. Thus, data were simulated using a compensatory two-dimensional IRT model in which the probability of a correct response is given as

$$P\left(X_{ij} = 1 \middle| \mathbf{a}_i, \mathbf{d}_i, \mathbf{\theta}_j\right) = \frac{\exp\left(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i\right)}{1.0 + \exp\left(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i\right)} , \qquad (12)$$

where

X_{ij} is the (0,1) score on item i by person j,

$\mathbf{a}_i = (a_{1i}, a_{2i}, ..., a_{ni})$ is the vector of item discrimination parameters,

d_i is a scalar difficulty parameter of item i, and

θ_{1j}, θ_{2j} is the two-dimensional trait parameter vector for person j.

This model is compensatory because of the addition of terms in the logit. This feature makes it possible for an examinee with low ability on one dimension to compensate by having a higher level on the remaining dimension(s) (Ackerman, 1989; Spray, Davey, Reckase, Ackerman, & Carlson, 1990).

Items in a two-dimensional latent space (e.g., math and verbal trait dimensions) can be conceptualized as vectors, following the work of Reckase (1985; Reckase & McKinley, 1991). Using Reckase's vector representation, a_{1i} and a_{2i} designate the composite of θ_1 and θ_2 measured by item i. If $a_{1i} = a_{2i}$, both traits would be measured equally well. However, if $a_{1i} = 0$ and $a_{2i} = 1.0$, discrimination would occur only along the θ_2 dimension with little or no discrimination among the levels of θ_1, depending on the correlation between θ_1 and θ_2. If all items in a test measure the same θ_1, θ_2 composite, the test would be considered unidimensional. In such an instance, only impact or true trait level differences on the valid composite of multiple traits could occur. The more varied the composites measured by the items in a test, the greater the violation of the assumption of unidimensionality and, hence, the greater the likelihood that DIF will occur.

Graphically, when items are represented as vectors, the length of the vector is equal to the amount of multidimensional discrimination, MDISC. For an item i this can be computed using

$$\text{MDISC} = \left(a_{1i}^2 + a_{2i}^2\right)^{\frac{1}{2}} \qquad (13)$$

MDISC is analogous to the unidimensional IRT model's discrimination parameter. For this study, level of discrimination was not a factor of interest. Thus, for all items the value of MDISC was fixed at 1.5.

The direction or θ_1, θ_2 composite being best measured is denoted by a reference angle that is given in degrees from the positive θ_1 axis and computed by

$$\alpha_i = \arccos\left[\frac{a_{1i}}{\text{MDISC}_i}\right]. \qquad (14)$$

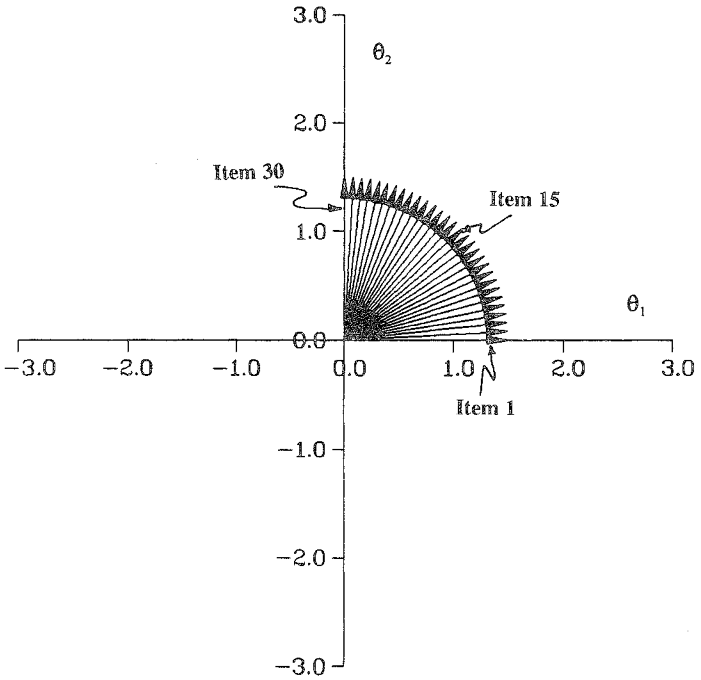
An item's vector originates at, and is graphed orthogonal to, the $p = .5$ equiprobability contour of the two-dimensional response surface. In the compensatory model these contours are always parallel.

Data Generation

Response data were generated for a 30-item test. The test consisted of items that had reference angles evenly spaced from 0° to 90° in approximately 3° increments. Specifically, for Item 1 (which measured only θ_1) $a_1 = 1.5$ and $a_2 = 0.0$; for Item 15 (which measured θ_1 and θ_2 approximately equally) $a_1 = 1.089$ and $a_2 = 1.032$; and for Item 30 (which measured only θ_2) $a_1 = 0.0$ and $a_2 = 1.5$. The difficulty parameter was set equal to 0.0 for all items. A plot of the 30 item vectors is shown in Figure 1.

The underlying trait distributions for the reference

Figure 1
An Item Vector Plot Illustrating the Angular Composites of the 30-Item Simulated Test



and focal groups were selected so that the reference group would have a higher mean θ_1 and the focal group a higher mean θ_2. Specifically, the $[\mu_{\theta_1}, \mu_{\theta_2}]$ vector for the reference group was [1.0, 0.0] and for the focal group, [0.0, 1.0]. For both groups, the θ_1 and θ_2 variances were set equal to 1.0 with a correlation between traits of .4 to reflect typical distributional values found in two-dimensional IRT estimates of achievement test data. The θ_1 and θ_2 values were randomly generated and restricted to a range of −2.5 to 2.5. Although these distributions were created for illustration purposes, they could realistically result from two groups being exposed to different instructional techniques within the same curriculum (e.g., one instructor emphasizes critical problem analysis, and another emphasizes computational algorithms only). The underlying trait distributions for each group along with their marginal trait distributions are shown in Figure 2.

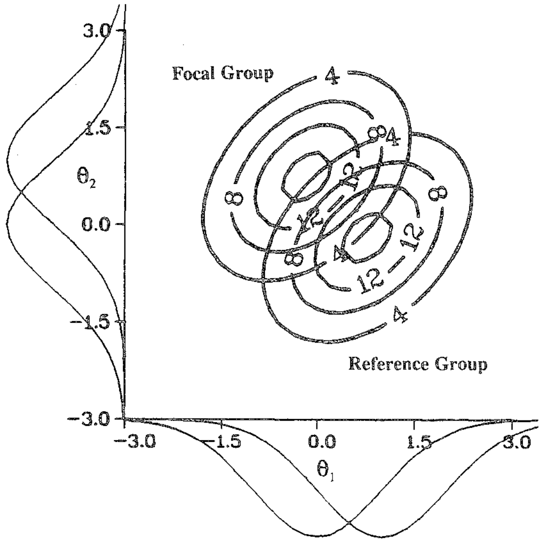
To study the DIF detection rates when the complete latent space is identified by the conditioning variable, three different reference group and focal group sample sizes were generated: $N_R = 1{,}000$, $N_F = 250$; $N_R = 1{,}000$, $N_F = 500$; and $N_R = 500$, $N_F = 250$. These values were selected to simulate realistic sample sizes.

For didactic purposes, this study was designed to imitate ideal (and perhaps unrealistic) conditions. That is, in addition to using the observed NC score as an estimate of an examinee's trait level for the conditioning variable, examinees also were matched on a linear transformation of their true latent traits. For each sample size, four separate DIF analyses (using both MH and SIB) were conducted.

Hypotheses and Analyses

Hypotheses. It is important to understand that the type of conditioning score used determines the valid test direction. Ackerman (1992) defined the two-dimensional sector that surrounds the valid test direction as the validity sector (i.e., the sector that contains the vectors representing the most valid

Figure 2
Contours and Marginal Distributions of the Generating Two-Dimensional Trait Distributions for the Reference and Focal Groups



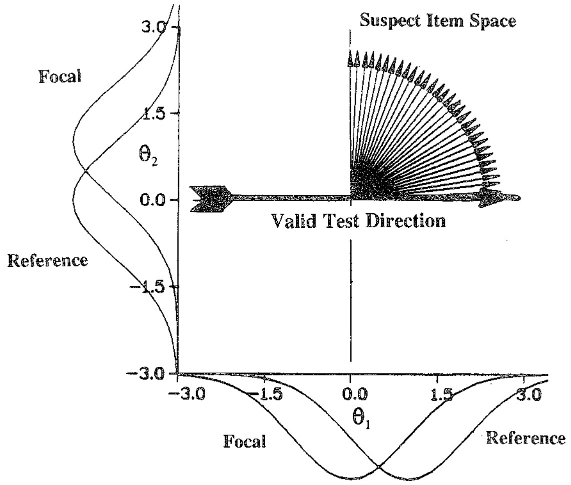
items). As the angular composite of the item departs from the valid test direction (i.e., begins to leave the validity sector), it becomes a suspect item or enters what is termed the suspect item space. Four different results were hypothesized, based on four different types of conditioning variables. Hypothesis 1 postulated that the NC score would correlate most highly with the linear θ_1, θ_2 composite that represented an equal weighting of both dimensions [i.e., X would correlate most highly with (cos 45°)θ_1 + (sin 45°)θ_2; hence, the valid test direction would be 45°]. Kim & Stout (1993) found a similar result when simulating a test that contained 20 items that measured only θ_1 and 20 items that measured only θ_2. As a result, when the NC score was used as the conditioning variable, there would be two suspect item spaces: one near the θ_1 axis and one near the θ_2 axis. That is, as the deviation of an item's angular composite direction from the valid direction increased, the more DIF (as measured by the size of the DIF statistic) would be exhibited in the item. Specifically, it was suspected that items measuring mostly θ_1 would show DIF against the focal group and items measuring mostly θ_2 would show DIF against the reference group.

Hypothesis 2 was that when a linear transformation of θ_1 was used as the conditioning variable (i.e., 0° would represent the valid test direction), there would be only one suspect item space: items measuring mostly θ_2 would show DIF against the reference group. Hypothesis 3 was the reverse of Hypothesis 2: When a linear transformation of θ_2 was used as the valid subtest score, items measuring mostly θ_1 would be identified as showing DIF against the focal group. For both Hypothesis 2 and Hypothesis 3, it was again believed that the degree of DIF would increase as the deviation of the items' angular composite increased from the valid test direction. A graphical illustration of Hypothesis 2 is illustrated in Figure 3. [Shealy & Stout (1993a, 1993b) and Ackerman (1992) provide a more detailed description of the conditional probability involving only θ_1 or only θ_2.]

Hypothesis 4 was that if the conditioning variable was both θ_1 and θ_2 (i.e., the complete latent space), none of the items would be identified as showing DIF against either group. In this situation there is no suspect item space because all items are considered to be valid.

Analyses. The four analyses paralleled the four

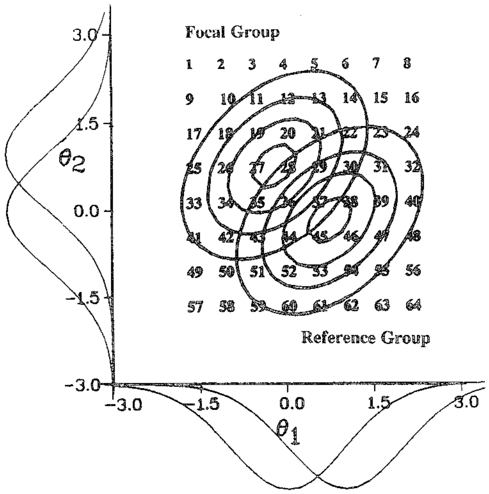
Figure 3
The Valid Test Direction and Suspect Item Space When Conditioning On a Linear Transformation of θ_1



hypotheses. In Analysis 1, the conditioning variable was the generated NC score. In Analysis 2, the examinee's θ_1 was transformed by X_1 = int(10θ_1) + 25, where "int" is the nearest integer of the value in the parentheses, and X_1 was used as the conditioning variable. In Analysis 3, the conditioning variable was a transformation of the examinee's θ_2 [i.e., X_2 = int(10θ_2) + 25]. In Analyses 2 and 3 there were 51 possible conditioning categories (0–50). These analyses were used to simulate situations in which DIF analyses are conducted on multidimensional tests, but the conditioning variable (e.g., the observed NC score) does not account for the entire latent trait space.

For Analysis 4, a conditioning score was used that accounted for the complete latent trait space; in this case, scores were conditioned on both θ_1 and θ_2. This was accomplished by arbitrarily imposing an 8 × 8 unit grid on the two-dimensional trait plane

Figure 4
Assignment of Conditioning Scores Using an 8 × 8 Grid



and assigning examinees within a cell a unique "score" from 1 to 64; these scores were used as the conditioning variable. The grid superimposed on the contours of the underlying trait distributions is illustrated in Figure 4.

Each of the four analyses was replicated 100 times. A different set of examinees was generated randomly each time from the specified underlying trait distributions. For each analysis, the mean (M) and standard deviation (SD) of MH and B_U were computed as well as the percent of times each item was statistically identified by each method ($\alpha = .05$) as biased against the reference group (P_R) and the percent of times it was identified ($\alpha = .05$) as biased against the focal group (P_F). Because MH does not indicate the direction of the DIF, the estimator ΔMH (Holland & Thayer, 1988, p. 135) was computed to identify the disadvantaged group.

The type of conditioning score used determines the valid test direction. Ackerman (1992) defined the two-dimensional sector that surrounds the valid test direction as the validity sector (i.e., the sector that contains the vectors representing the most valid items). As the angular composite of the item departs from the valid test direction (i.e., begins to leave the validity sector), it becomes a suspect item or enters what is termed the suspect item space.

Results

The results are summarized in Tables 2–5 (one for each hypothesis/analysis). Because of the high degree of similarity between all sample size conditions, only the results for the smallest sample size condition—$N_R = 500, N_F = 250$—are presented in the tables. The only difference across sample sizes was that both MH and B_U detected slightly more items as consistently showing DIF against one group or the other as sample sizes increased.

Analysis 1

Table 2 displays the results when the NC score was used as the conditioning variable. As predicted, for each sample size the NC score correlated most highly with the linear θ_1, θ_2 composite that represented an equal weighting of both dimensions (as compared with all other θ_1, θ_2 composites). The correlations were .78 for the $N_R = 1{,}000, N_F = 250$ condition; .73 for $N_R = 1{,}000, N_F = 500$; and .69 for $N_R = 500, N_F = 250$. Also as predicted, items measuring mostly θ_1 (Items 1–14) or mostly θ_2 (Items 16–30; see Figure 1) were identified as exhibiting DIF. In the $N_R = 500, N_F = 250$ condition, Items 1–5 and 24–30 were identified 100% of the time as significantly favoring the reference group or the focal group, respectively. In the $N_R = 1{,}000, N_F = 250$ condition, Items 1–6 and 24–30 were identified 100% of the time; and in the $N_R = 1{,}000, N_F = 500$ condition, Items 1–9 and 22–30 were identified 100% of the time. MH and B_U performed equally well.

Analysis 2

When conditioning on a linear transformation of θ_1, items with angular composites greater than 30° (Items 12–30) were consistently identified by both DIF statistics as showing DIF against the reference group. These results are shown in Table 3. For the $N_R = 1{,}000, N_F = 250$ condition, Items 11–30 were rejected 100% of the time; for the $N_R = 1{,}000, N_F = 500$ condition, Items 9–30 were rejected 100% of the time. Slight differences were noted between MH and B_U. B_U seemed to be more sensitive (the rejection rate for B_U increased at a faster rate than MH as the angular composite of the item departed from 0°). This is shown by the fact that P_R is .9 or above by Item 7 for B_U but not until Item 9 for MH. This was expected because for Hypotheses 2–4 the regression correction for B_U was intentionally not used because the conditioning variable was based on a latent trait parameter rather than a test score; therefore, using the regression correction was inappropriate. Without the regression correction, greater rejection rates are expected for B_U. Thus, MH and B_U should not be directly compared based on data in Table 3.

Analysis 3

As hypothesized, the opposite results occurred when the valid test direction was along the θ_2 axis, as shown in Table 4. When a linear transformation of θ_2 was used as the conditioning criterion, items with angular composites less than 60° (Items 1–18) were consistently identified by MH and B_U as significantly favoring the reference group for the $N_R =$

Table 2
M, SD, P_R, and P_F for MH and B_U When the Conditioning Variable Was the NC Score (Analysis 1) for $N_R = 500, N_F = 250$

	MH				B_U			
Item	M	SD	P_R	P_F	M	SD	P_R	P_F
1	49.14	13.78	0.00	1.00	−7.54	1.23	0.00	1.00
2	42.28	11.74	0.00	1.00	−7.03	1.12	0.00	1.00
3	37.26	11.08	0.00	1.00	−6.54	1.07	0.00	1.00
4	34.06	10.91	0.00	1.00	−6.25	1.10	0.00	1.00
5	30.22	11.51	0.00	1.00	−5.82	1.11	0.00	1.00
6	25.99	9.83	0.00	.99	−5.42	1.13	0.00	.99
7	20.10	7.96	0.00	1.00	−4.75	1.02	0.00	1.00
8	17.54	8.89	0.00	.99	−4.39	1.16	0.00	.99
9	12.42	6.20	0.00	.93	−3.69	.99	0.00	.95
10	8.81	5.57	0.00	.79	−3.11	1.05	0.00	.86
11	6.15	4.14	0.00	.68	−2.59	.93	0.00	.73
12	4.46	3.64	0.00	.51	−2.07	1.08	0.00	.57
13	2.68	3.22	0.00	.27	−1.48	1.05	0.00	.32
14	1.63	2.06	0.00	.13	−1.00	1.05	0.00	.22
15	.86	1.24	.01	.02	−.30	1.03	.01	.06
16	.84	1.14	.02	.01	.30	1.07	.08	.02
17	1.58	2.01	.11	0.00	.96	1.12	.16	.02
18	2.91	2.97	.29	0.00	1.60	1.09	.36	0.00
19	4.03	3.70	.36	0.00	2.05	1.05	.54	0.00
20	7.14	5.43	.68	0.00	2.84	1.15	.78	0.00
21	9.19	6.01	.79	0.00	3.25	1.18	.87	0.00
22	12.38	6.12	.94	0.00	3.88	1.05	.96	0.00
23	15.88	7.78	.96	0.00	4.39	1.16	.97	0.00
24	20.50	7.75	1.00	0.00	5.12	1.09	1.00	0.00
25	23.45	9.20	1.00	0.00	5.50	1.24	1.00	0.00
26	28.28	10.06	1.00	0.00	6.05	1.22	1.00	0.00
27	31.19	9.20	1.00	0.00	6.34	1.09	1.00	0.00
28	34.97	10.21	1.00	0.00	6.77	1.15	1.00	0.00
29	42.19	13.29	1.00	0.00	7.46	1.41	1.00	0.00
30	46.74	11.52	1.00	0.00	7.96	1.17	1.00	0.00

500, $N_F = 250$ condition. For the $N_R = 1{,}000$, $N_F = 250$ condition, Items 1–20 were always identified; for the $N_R = 1{,}000$, $N_F = 500$ condition, Items 1–23 were consistently identified. Thus, as the sample size increased, the sensitivity of the statistics to depart slightly from the valid test direction also increased. Again B_U had higher rejection rates than did the MH statistic because the regression correction was not used.

Analysis 4

For Analysis 4, the conditioning variable was used to match examinees on their complete vector of θs used to generate the response data. The results are given in Table 5. Unlike the previous analyses in which examinees were matched on only a single trait or composite trait, neither the mean of the MH or the mean of B_U were statistically significant for any of the items for any of the three sample size conditions. For the reasons discussed above, B_U produced greater rejection rates (i.e., Type I error because no DIF occurred) for each of the sample size combinations.

Summary

Plots of the mean B_U and mean MH results for the $N_R = 500$, $N_F = 250$ condition are presented in Figures 5a and 5b, respectively, for each item and for each conditioning score. Regions for which items would significantly favor the reference group

Table 3
M, SD, P_R, and P_F for MH and B_U When the Conditioning Variable Was a Transformation of θ_1 (Analysis 2) for $N_R = 500$, $N_F = 250$

	MH				B_U			
Item	M	SD	P_R	P_F	M	SD	P_R	P_F
1	.90	1.37	.01	.01	.55	1.56	.21	.03
2	.95	1.32	.07	0.00	1.31	1.35	.30	.01
3	1.54	2.12	.13	0.00	1.87	1.35	.43	0.00
4	2.55	2.62	.21	0.00	2.58	1.29	.68	0.00
5	3.68	3.55	.39	0.00	2.97	1.50	.77	0.00
6	5.02	4.42	.50	0.00	3.49	1.75	.82	0.00
7	7.65	4.90	.74	0.00	4.38	1.55	.93	0.00
8	9.03	5.93	.86	0.00	4.70	1.77	.95	0.00
9	11.67	6.11	.93	0.00	5.30	1.70	.97	0.00
10	14.89	6.34	.97	0.00	5.99	1.36	1.00	0.00
11	16.76	5.89	.99	0.00	6.40	1.59	1.00	0.00
12	18.89	7.17	1.00	0.00	6.66	1.62	.99	0.00
13	22.17	7.85	1.00	0.00	7.36	1.61	1.00	0.00
14	23.72	7.74	1.00	0.00	7.58	1.65	1.00	0.00
15	26.15	8.82	1.00	0.00	8.09	1.79	1.00	0.00
16	28.32	8.67	1.00	0.00	8.27	1.57	1.00	0.00
17	31.15	10.01	1.00	0.00	8.60	1.88	1.00	0.00
18	35.32	10.98	1.00	0.00	9.26	2.01	1.00	0.00
19	35.23	9.79	1.00	0.00	9.27	1.70	1.00	0.00
20	38.41	10.51	1.00	0.00	9.63	1.85	1.00	0.00
21	40.76	11.34	1.00	0.00	9.94	1.83	1.00	0.00
22	42.88	10.37	1.00	0.00	10.02	1.64	1.00	0.00
23	43.76	11.78	1.00	0.00	10.13	1.87	1.00	0.00
24	45.78	11.18	1.00	0.00	10.26	1.74	1.00	0.00
25	47.61	11.16	1.00	0.00	10.48	1.82	1.00	0.00
26	48.45	11.38	1.00	0.00	10.31	1.67	1.00	0.00
27	48.63	11.81	1.00	0.00	10.34	1.70	1.00	0.00
28	48.26	11.28	1.00	0.00	10.22	1.74	1.00	0.00
29	51.32	13.49	1.00	0.00	10.57	1.82	1.00	0.00
30	53.17	13.13	1.00	0.00	10.54	1.82	1.00	0.00

or significantly favor the focal group are indicated. Both plots show that the conditioning score used in Analysis 4 (θ_1, θ_2) eliminated the DIF from all items.

Discussion

The purpose of this paper was not to compare the MH and SIB procedures. Rather, it was to demonstrate what causes DIF and how different conditioning scores yield different results. All of the obtained results were predicted from theory. If the complete latent trait space could be used as the conditioning variable, DIF would be eliminated. However, conditioning on the complete latent space should not be the intent of DIF analyses. Testing practitioners should instead condition on the score that best identifies the trait intended to be measured by the test. Subsequently, any items that display DIF must be studied further.

Examinee-item interaction is never simple and rarely predictable, especially for large populations. Despite the care and caution that is taken in writing and reviewing items before they are used in a test, some items are sensitive to differences in extraneous traits. These are the items that must be identified and studied. However, the success of identifying these problematic items depends on how well those items that are measuring only the intended trait can be identified. Shealy & Stout (1993a, 1993b) urged practitioners to identify the most valid items and condition on the NC score obtained for these items. Ackerman

Table 4
M, SD, P_R, and P_F for MH and B_U When the Conditioning Variable Was a Transformation of θ_2 (Analysis 3) for $N_R = 500$, $N_F = 250$

	MH				B_U			
Item	M	SD	P_R	P_F	M	SD	P_R	P_F
1	58.11	15.00	0.00	1.00	−9.03	1.45	0.00	1.00
2	57.52	14.36	0.00	1.00	−8.93	1.38	0.00	1.00
3	56.66	14.77	0.00	1.00	−8.91	1.46	0.00	1.00
4	56.45	13.54	0.00	1.00	−8.81	1.31	0.00	1.00
5	54.80	14.86	0.00	1.00	−8.57	1.45	0.00	1.00
6	54.37	13.56	0.00	1.00	−8.50	1.23	0.00	1.00
7	50.19	12.79	0.00	1.00	−8.14	1.20	0.00	1.00
8	50.84	13.90	0.00	1.00	−8.08	1.35	0.00	1.00
9	46.75	12.73	0.00	1.00	−7.68	1.33	0.00	1.00
10	44.99	12.25	0.00	1.00	−7.43	1.21	0.00	1.00
11	42.31	10.25	0.00	1.00	−7.19	1.14	0.00	1.00
12	41.11	12.09	0.00	1.00	−6.97	1.28	0.00	1.00
13	37.19	12.23	0.00	1.00	−6.62	1.30	0.00	1.00
14	35.21	11.66	0.00	1.00	−6.42	1.29	0.00	1.00
15	32.07	10.21	0.00	1.00	−6.02	1.16	0.00	1.00
16	29.01	10.12	0.00	1.00	−5.68	1.21	0.00	1.00
17	25.66	10.87	0.00	1.00	−5.15	1.30	0.00	.99
18	22.96	8.68	0.00	1.00	−4.91	1.02	0.00	1.00
19	21.83	9.21	0.00	.98	−4.84	1.12	0.00	.99
20	17.47	8.36	0.00	.99	−4.20	1.10	0.00	.99
21	14.70	7.52	0.00	.96	−3.85	1.13	0.00	.97
22	11.26	5.63	0.00	.93	−3.39	.94	0.00	.96
23	9.90	6.12	0.00	.86	−3.06	1.06	0.00	.87
24	7.26	4.82	0.00	.69	−2.60	1.05	0.00	.72
25	5.93	4.51	0.00	.57	−2.32	1.07	0.00	.57
26	3.73	3.37	0.00	.40	−1.81	1.10	0.00	.39
27	2.88	3.02	0.00	.31	−1.52	1.09	0.00	.39
28	1.77	2.15	0.00	.13	−1.14	.89	0.00	.21
29	1.31	1.74	.01	.08	−.54	1.22	.02	.13
30	.76	1.14	.01	.02	.14	1.10	.05	.03

(1992) suggested that one approach would be to use both multidimensional IRT and a thorough review of the items to try to identify sectors that contain the most valid items.

Determining the variables or scores on which to condition to account for the complete latent trait space is obviously not an easy task. Any linear combination of item scores does not remove DIF—it simply redefines the valid test direction, which in turn determines which items will exhibit DIF. As mentioned above, this study represented the ideal case in which the underlying traits were known.

It is doubtful that practitioners will ever be able to account for the complete latent space, or be able to condition on scores that can account for the complete latent space. It would, however, be useful to identify scores or variables that decrease the amount of DIF, because it is these variables that will provide insight into what the items are actually measuring. The search for these variables is part of the construct validation process. Finding them will also help to explain why the DIF occurred.

Differences between the performance of MH and B_U, especially for Analysis 4 (in which the conditioning score accounted for the complete latent trait space), may be due in part to the way the statistics were computed in this study. To simplify the process, the iterative purification process in which the MH and SIB analyses are typically conducted examining one item at a time was not used. Thus, for

Table 5
M, SD, P_R, and P_F for MH and B_U When the Conditioning Variable Was θ_1 and θ_2 (Analysis 4) for $N_R = 500, N_F = 250$

	MH				B_U			
Item	M	SD	P_R	P_F	M	SD	P_R	P_F
1	.95	1.28	0.00	.06	−.24	2.24	.18	.22
2	.74	1.03	0.00	.02	−.44	2.26	.15	.28
3	.72	1.05	.01	.01	−.32	1.91	.13	.22
4	.64	.91	0.00	.03	−.56	1.83	.13	.23
5	.91	1.34	.01	.05	−.26	1.99	.11	.19
6	.98	1.71	0.00	.05	−.40	1.87	.12	.18
7	.61	.93	0.00	.01	−.22	1.89	.10	.14
8	.97	1.47	0.00	.07	−.35	2.20	.18	.26
9	.83	1.13	0.00	.02	−.23	1.99	.14	.21
10	.70	1.12	.02	.02	−.12	2.02	.14	.14
11	.63	1.20	0.00	.01	−.22	2.08	.17	.20
12	.73	.98	0.00	.01	−.42	1.91	.08	.24
13	.66	1.10	.02	0.00	−.20	1.91	.15	.20
14	.84	1.39	.03	.01	−.17	2.15	.16	.20
15	.70	1.00	.01	0.00	−.12	2.06	.14	.15
16	.80	1.41	0.00	.04	−.26	1.94	.12	.19
17	.86	1.25	.02	.03	.11	2.20	.17	.15
18	.93	1.49	.02	.02	−.05	2.52	.21	.22
19	.71	1.19	.02	0.00	−.24	1.77	.11	.13
20	.68	.92	0.00	.01	.09	2.02	.16	.12
21	1.07	1.79	.05	.01	.12	2.14	.20	.18
22	.84	1.18	.03	.01	.44	1.87	.19	.08
23	.83	1.27	.03	.02	.50	1.98	.22	.11
24	.76	1.11	.03	.01	.19	1.83	.15	.13
25	.83	1.41	.03	.01	.17	1.96	.20	.10
26	.73	1.19	.03	.01	.05	1.90	.14	.15
27	.69	.91	.01	0.00	.19	1.88	.17	.13
28	.81	1.20	.02	0.00	.01	1.80	.17	.14
29	1.19	1.45	.05	0.00	.46	2.03	.21	.09
30	.99	1.51	.04	.01	.37	1.79	.22	.08

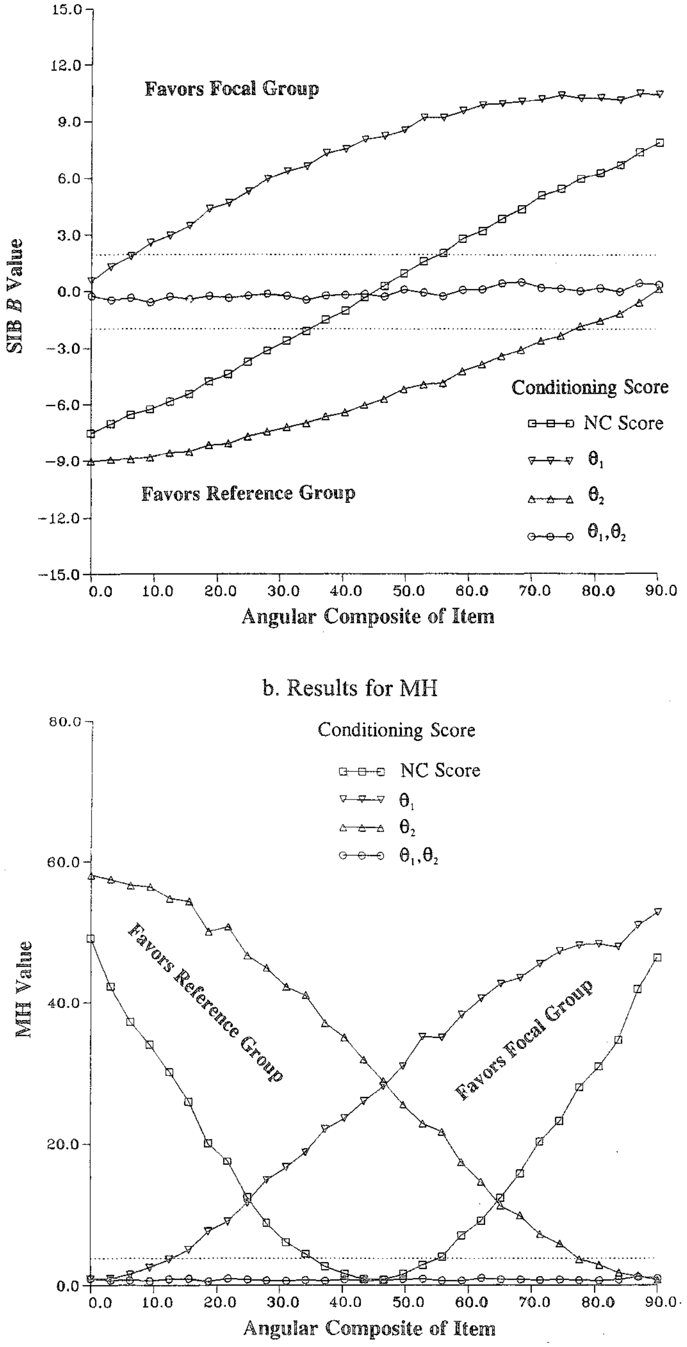
Analyses 1–3, the scores always contained the influence of DIF items other than just the studied item.

To control for impact, the SIB statistic requires a regression correction. This correction was used only when the conditioning score was number correct (Analysis 1). It was inappropriate when the conditioning variable was a transformation of θ_1, θ_2, as discussed above. For Hypothesis 4, when conditioning was achieved by imposing a coarse 8 × 8 grid, scores were a result of the approximate (θ_1, θ_2) location in the latent trait plane and were not a monotone function of the two trait parameters. The result of not applying the regression correction in this case may have caused the higher rates of rejection for SIB compared to those for the MH statistic. That is, the size of each of the 8 × 8 cells may have been large enough to permit distributional differences between the reference and focal groups to occur, confounding DIF with impact. This was particularly noticeable for the coarse 8 × 8 grid conditioning for Hypothesis 4. Although unplanned, this result was important because it demonstrated the importance of the regression correction: Even though examinees within a relatively small area on the θ_1, θ_2 plane were assigned the same score, there was a potential for confounding DIF with impact.

In the analyses for Hypotheses 1, 2, and 3 the size or amount of DIF increased as the item's angular composite direction departed from the valid test direction. For large sample sizes (i.e., the $N_R = 1{,}000$,

Figure 5
Mean DIF Estimates Over 100 Replications for Each Item and for Each Type of Conditioning Score ($N_R = 500$, $N_F = 250$ Sample Size)

a. Results for B_U



b. Results for MH

$N_F = 500$ condition), the power of the MH and SIB procedures to detect DIF increased and, thus, the angular difference between the valid test direction and the optimal measurement direction of an item needed to consistently achieve statistical significance can be quite small (i.e., less than 30°).

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113–127.

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. *Journal of Educational Measurement, 24,* 67–91.

Ackerman, T. A., & Evans, J. A. (1992, April). *An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In H. Wainer & P. W. Holland (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.

Kim, H. R., & Stout, W. F. (1993, April). *A robustness study of ability estimation in the presence of latent trait multidimensionality using the Junker/Stout index ε of dimensionality.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta GA.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

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

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

Roussos, L. A., Stout, W. F., & Marden, J. I. (1993, April). *Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis.* Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta GA.

Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239) Hillsdale NJ: Erlbaum.

Shealy, R., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Shin, S. (1992). An empirical investigation of the robustness of the Mantel-Haenszel procedure and sources of differential item functioning. *Dissertation Abstracts International, 53A,* 3504.

Spray, J. A., Davey, T., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (ACT Research Report—ONR 90-8). Iowa City IA: American College Testing Program.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. *Journal of Educational Measurement, 26,* 55–66.

Author's Address

Send requests for reprints or further information to Terry Ackerman, Department of Educational Psychology, University of Illinois, Champaign IL 61820-6990, U.S.A. Internet: ackerman@cso.vmd.uiuc.edu.