ED 359 223 TM 019 986

AUTHOR Ackerman, Terry A.; Evans, John A.

TITLE A Didactic Example of the Influence of Conditioning on the Complete Latent Ability Space When Performing DIF Analyses.

PUB DATE Apr 93

NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Computer Simulation; Equations (Mathematics); *Item Bias; Item Response Theory; *Mathematical Models; *Predictive Measurement; Sample Size; *Test Items

IDENTIFIERS *Latent Space; *Mantel Haenszel Procedure; Simultaneous Item Bias Procedure

ABSTRACT

A didactic example is provided, using a Monte Carlo method, of how differential item functioning (DIF) can be eliminated (and thus better understood) when the complete latent space is used. The main source of DIF is that the matching single criterion used in some DIF procedures, Mantel Haenszel or Simultaneous Item Bias (SIBTEST), does not account for all of the abilities used by examinees in both groups of interest. To resolve this problem, several researchers have tried to match on secondary variables, but with no success. In this paper, response data are generated from a 2-dimensional item response theory model for a 30-item test in which items are measuring uniformly spaced composites of theta(sub 1) and theta(sub 2). Two DIF detection methods, the Mantel Haenszel and SIBTEST detection procedures, are used under 3 reference/focal sample size conditions (1,000/250, 1,000/500, and 500/250). When the procedures conditioned on the number correct score, only on theta(sub 1), or only on theta(sub 2), the differential group performance followed predictable patterns. Likewise, when the matching criterion was a function of both thetas (i.e., the complete latent space was identified), the DIF was eliminated for all items as hypothesized. Twelve tables and two figures illustrate the analyses. (SLD)

# A Didactic Example of the Influence of Conditioning on the Complete Latent Ability Space When Performing DIF Analyses

Terry A. Ackerman
University of Illinois

John A. Evans
Peoria Public School System

Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, Georgia, April, 1993.

# A Didactic Example of the Influence of Conditioning on the Complete Latent Ability Space When Performing DIF Analyses

## Abstract

The purpose of this paper is to provide a didactic example of how differential item functioning (DIF) can be eliminated (and thus better understood) when the complete latent space is used. The main source of DIF is that the matching single criterion used in some DIF procedures (e.g., Mantel Haenszel or SIBTEST) does not account for all of the abilities used by examinees in both groups of interest. To resolve this problem several researchers have tried to match on secondary variables, but with no success. In this paper response data is generated from a two-dimensional item response theory model for a 30-item test in which items are measuring uniformly spaced composites oı $\theta_1$ and $\theta_2$. Two different DIF detection methods, the Mantel Haenszel and Simultaneous Item Bias (SIBTEST) detection procedure are used under three different sample size conditions. When the procedures conditioned on the number correct score, or only on $\theta_1$, or only on $\theta_2$, the differential group performance followed predictable patterns. Likewise when the matching criterion was a function of both $\theta_1$ and $\theta_2$ (i.e., the complete latent space was identified) the DIF was eliminated for all items as hypothesized.

*Introduction*

The main source of differential item functioning (DIF) is that the matching criteria does not account for the complete latent space of abilities that was used by the examinees in both groups of interest. Specifically, according to the bias/DIF theory of Shealy and Stout (in press), DIF or item bias occurs when items are capable of discriminating between levels of several abilities, the groups of interest have different distributions of these abilities, and the test results are summarized as only a single score (cf. Ackerman, 1992). When subjects are being measured on a single skill they can easily be ordered. However, if multiple skills are being assessed, but only a single score is reported, then differences in underlying two-dimensional ability distributions may cause individuals who have the same latent ability to receive quite different scores. To circumvent this problem researchers (Zwick & Ercikan, 1989; Shin, 1992) have tried, albeit unsuccessfully, to better account for the latent ability space by conditioning upon several variables.

The impetus for this study arose after considering these unsuccessful attempts by Zwick and Erickan and by Shin. In both of their studies efforts to condition upon a second variable to better account for differential performance, thereby diminishing the amount of bias present, did not prove successful. These researchers tried to remove the item performance differences between two groups of interest by matching subjects on two variables logically related to the underlying response process of the test.

The purpose of this research study is to demonstrate, using a monte carlo format, that if the complete latent ability space could be identified, differences in item performance between two groups of interest would be eliminated. Unlike the studies to date which have failed in their attempts to identify the complete latent ability space, working with generated multidimensional data has the advantage of knowing all of the true abilities which produced the observed responses. Two DIF approaches, the Mantel Haenszel procedure (Holland & Thayer, 1988) and the Simultaneous Item Bias (SIBTEST) detection procedure (Shealy & Stout, in press) are used under three different sample size conditions.

*Bias detection methodology*

Although there has been a proliferation of methods to detect item bias, this paper will focus on only two: the Mantel-Haenzsel (MH) procedure (Holland & Thayer, 1988), and Shealy and Stout's SIBTEST (in press). Both of these procedures are nonparametric and require no model calibration. However, they can be studied from an IRT framework, and as such can be explained within the IRT context already developed.

To compute the Mantel Haenszel statistic for an item *i* examinees from a reference and focal group are matched according to their number correct score. For each possible score category a 2 X 2 contingency table is created in which the frequency of correct and incorrect answers for each group are noted along with the marginal and total frequencies. The table for the *jth*-score category would have the following values:

| | Item Score | | |
|---|---|---|---|
| | 1 | 0 | |
| Reference | $A_j$ | $B_j$ | $n_{rj}$ |
| Fucal | $C_j$ | $D_j$ | $n_{fj}$ |
| | $m_{1j}$ | $m_{0j}$ | $T_j$ |

Summing over the contingency tables and using a continuity correction, the MH statistic is given by

$$M\hat{H}_j = \frac{(|\sum_j A_j - \sum_j E(A_j)| - \frac{1}{2})^2}{\sum_j Var(A_j)} \qquad (1)$$

where

$$E(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j}$$

and

$$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

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

$\hat{MH}$ is approximately distributed as $\chi^2$ with one degree of freedom when $H_o$ below holds and thus, can be used to test the null hypothesis that for each score category $j$ the odds of a reference group examinee getting the item correct equals the odds that a focal group examinee will get the item correct. Specifically, $H_o$: $\frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}} \quad j=1,...,k$ is tested against the alternative of uniform DIF; that is, $H_A$: $\frac{p_{Rj}}{q_{Rj}} = \alpha \frac{p_{Fj}}{q_{Fj}} \;, \quad \alpha \neq 1 \qquad j=1,...,k$.

The Shealy and Stout SIBTEST statistic, $\mathbf{B}_{uni}$, is computed in a manner somewhat similar to the Standardization procedure (Dorans & Kulick, 1986) with several important differences. First, its computation requires that the practitioner identify a *valid subtest* of items (possibly all of the items except the one item suspected of DIF) and a set of *suspect* item(s). For example, the valid items can be identified as the items which weight most highly on a particular factor of a factor analysis of the item tetrachoric matrix or via a hierarchical cluster analysis (Roussos, Stout, Marden, 1993) or by cognitive considerations. (Note: Shealy and Stout's valid subtest is somewhat comparable to the collection of non-DIF items identified using a purification approach with the MH procedure.) 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), $Y = \sum_{i=n+1}^{N} U_i$, and the valid subtest score, $X = \sum_{i=1}^{n} U_i$, are computed. $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ representing the average $Y$ for all examinees attaining a valid subtest score X = $k$ *(k=0,1,2..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 employed to attain adjusted values $\bar{Y}^*_{Rk}$ and $\bar{Y}^*_{Fk}$.

Shealy and Stout define ꝛ model based parameter $\beta_{uni}$ measuring the amount of unidirectional (non-crossing) bias present. An estimate $\hat{\beta}_{uni}$ of $B_{uni}$ is then defined as

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{p}_k (\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk})$$

where $\hat{p}_k = \frac{(J_{Rk} + J_{Fk})}{\sum_{j=0}^{n} (J_{Rj} + J_{Fj})}$ with $J_{Rk}$ and $J_{Fk}$ the number of examinees in the reference and focal groups with the same valid score X = *k*. The test statistic is given by

$$B_{uni} = \frac{\hat{\beta}_{uni}}{\hat{\sigma}(\hat{\beta}_{uni})}$$

where the denominator is the estimated standard error of $\hat{\beta}_{uni}$ and is computed as

$$\hat{\sigma}(\hat{\beta}_{uni}) = \left[ \sum_{k=0}^{n} \hat{p}_k^2 \left( \frac{1}{J_{Rk}} \hat{\sigma}^2(Y|k,R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y|k,F) \right) \right]^{\frac{1}{2}}$$

where the $\hat{\sigma}^2{}_S$ are the empirical cell *k* variances for the suspect test scores.

The test statistic has an approximate *N(0,1)* distribution when no DIF is present (i.e., $B_{uni} = 0$. Thus, the hypothesis of testing bias against the focal group can be stated as

$$H_o: \beta_{uni} = 0 \qquad vs \qquad H_A: \beta_{uni} > 0$$

### Method

The simulation of DIF in this study is done using a two-dimensional IRT model because it is believed to be a more valid approach than using a unidimensional IRT model and assigning different generating item parameters to each group depending on the direction of bias. This is consistent with the Shealy and Stout (1993) theory of test bias. DIF is caused by the inability of a single score or a unidimensional latent ability estimate to account for the entire latent ability space. Ackerman (1992) outlined ways in which the underlying ability distributions could produce differences in rescaled unidimensional item characteristic curves 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

$$P(X_{ij} = 1 | a_i, d_i, \theta_j) = \frac{e^{a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i}}{1.0 + e^{a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i}}$$

where $X_{ij}$ is the score (0,1) on item *i* by person *j*, $\boldsymbol{a}_i$ is the vector of item discrimination parameters, $d_i$ is a scalar difficulty parameter of item *i*, and $(\theta_{1j},\theta_{2j})$ is the vector of ability parameters for person *j*.

Perhaps the best way to think of items in a two-dimensional latent space (e.g., math and verbal ability dimensions) is to conceive of them as vectors following the work of Reckase (1986). Utilizing Reckase's vector representation the $a_{1i}$ and $a_{2i}$ discrimination parameters designate the composite of $\theta_1$ and $\theta_2$ that item *i* is measuring. If $a_{1i} = a_{2i}$ both abilities would be measured equally well. However, if $a_{1i} = 0$ and $a_{2i} = 1.0$, discrimination would only occur along the $\theta_2$ dimension with little or no discrimination among the levels of $\theta_1$ depending on the correlation between $\theta_1$ and $\theta_2$. If all the items in a test are measuring exactly the same $(\theta_1, \theta_2)$ composite the test would be considered to be unidimensional. (Note, in such an instance only impact (true ability differences) could occur.) The more varied the composites a test is measuring, the more multidimensional the test and the greater the likelihood of DIF occurring.

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 the formula

$$MDISC = \sqrt{a_{1i}^2 + a_{2i}^2}$$

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

The direction or $(\theta_1,\theta_2)$ composite being best measured is denoted by a reference angle that is given in degrees from the positive $\theta_1$ axis and computed using the formula

$$\alpha_i - \arccos\left[\frac{a_{1i}}{MDISC_i}\right]$$

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.

For the purposes of this study response data was generated for a 30 item test. The test was highly multidimensional with items having reference angles from 0 to 90 degrees in approximately three degree increments. Specifically, the $(a_1, a_2)$ values for Item 1 (which measured only $\theta_1$) were (1.5, 0), for Item 15 (that measured $\theta_1$ and $\theta_2$ about equally), (1.089, 1.032) and for Item 30 (which measured only $\theta_2$), (0,1.5). The difficulty parameter was set equal to zero for all items. A plot of the 30 item vectors is shown in Figure 1.

Insert Figure 1 about here

The underlying ability distributions for the Reference and Focal groups were chosen so that the Reference group would have a higher $\theta_1$ mean ability and the Focal group a higher $\theta_2$ mean ability. Hence, 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 $\theta_1$ and $\theta_2$ variances were set equal to 1.0 with a correlation between abilities equal to .4. The $\theta_1$ and $\theta_2$ values were randomly generated and restricted to a range from -2.5 to 2.5. These distributions were created for illustration purposes, although they could realistically result from two groups being exposed to different instructional techniques within the same curriculum (e.g., one instructor emphasizing critical problem analysis, while another instructor emphasizes computational algorithms only). The underlying ability distributions for each group along with their marginal ability distributions are shown in Figure 2.

Insert Figure 2 about here

To study the power of SIBTEST and MH when the complete latent space is identified by the conditioning variable, three different Reference/Focal sample sizes were generated, 1000/250, 1000/500 and 500/250. These values were chosen to simulate realistic sample sizes that testing practitioners are likely to often work with.

For didactic purposes this study was set up to imitate ideal (and perhaps unrealistic) conditions. That is, instead of using the observed number correct score as an estimate of an examinee's ability for the conditioning variable, subjects were matched on a linear transformation of their true latent ability. How closely practitioners can come to duplicating this is a function of the test reliability. For each sample size, four separate DIF analyses (using both MH and SIBTEST) were conducted. In the first analysis, the matching variable was the generated number correct score. In the second analysis a transformation of the examinee's $\theta_1$ ability (i.e., $X_1 = \text{Int}(10\theta_1) + 25$ ), where "Int" represents the nearest integer of the value in the parenthesis. In the third analysis the conditioning variable was a transformation of only the examinee's $\theta_2$ ability (i.e., $X_2 = \text{Int}\ (10\theta_2) + 25$). In both of the second and third analyses there were 51 possible matching categories, (0 - 50). These analyses were used to simulate scenarios in which DIF analyses are conducted on multidimensional tests, but the matching variable (e.g., the observed number correct score) does not account for the entire latent ability space.

The final analysis depicted a situation in which the DIF analysis was able to be matched on either a single or multiple variable(s) that represented all the abilities that were used in the response process conditioning on both $\theta_1$ and $\theta_2$. This was carried out by imposing a 8 X 8 unit grid on the two-dimensional ability plane and each cell of this grid was assigned a theoretical "score" from 1 to 64. Thus, in this analysis there were 64 possible score categories to condition upon. Each of the four analyses were replicated 100 times, each time with a new set of examinees from the specified underlying ability distributions. For each analysis the mean, standard deviation, the percent of times an item was statistically flagged ($\alpha = .05$) as favoring the reference group and the percent of times it was statistically flagged ($\alpha = .05$) as favoring the focal group were computed for both DIF statistics. To determine the direction of the bias for the MH results, the estimator $\Delta\hat{MH}$ (Holland and Thayer, 1988) was computed.

It is important to understand that the type of conditioning score employed 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 which contains the vectors representing the most valid items). As the angular composite of the item begins to depart from the valid test direction (i.e., begins to leave the validity sector) it then becomes a suspect item or enters what is tenned the *suspect item space*.

In this study four different results were hypothesized. It was postulated that the number correct score would correlate most highly with the linear $\theta_1,\theta_2$ composite that represented an equal weighting of both dimensions (i.e., X would correlate most highly with (cos 45°)$\theta_1$ + (sin 45°)$\theta_2$ ; hence, the valid test direction would be 45°). Kim (1993) found a similar result when simulating a test that contained 20 items that measured only $\theta_1$ and 20 items that measured only $\theta_2$. As a result, when the number correct score was used as the conditioning variable there would be two suspect item spaces: one near the $\theta_1$-axis and one near the $\theta_2$-axis. That is, as the deviation of an items' angular composite direction from the valid direction increased, the more biased (as measured by the size of the bias statistic) the item would become. Specifically, it was suspected that items which were measuring mostly $\theta_1$ would be biased against the focal group and items that tended to measure mostly $\theta_2$ would be biased against the reference group.

The second hypothesis was that when a linear transformation of $\theta_1$ 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 $\theta_2$ would be biased against the reference group. The third hypothesis was the reverse of the second: when a linear transformation of $\theta_2$ was used as the valid subtest score, items measuring mostly $\theta_1$ would be flagged as being biased against the focal group. For both hypothesis two and three, it was again believed that the degree of bias would increase as the deviation of the items' angular composite increased from the valid test direction. The final hypothesis was that if the conditioning variable was jointly $\theta_1$ and $\theta_2$ (i.e., the complete latent space) none of the items would be flagged as being biased against either group. In this situation there is no suspect item space.

## Results

The results of the simulation runs are summarized in Tables 1-12 (four tables for each sample size). All of the hypotheses were substantiated.

*Hypothesis One:*

Tables 1 - 3 display the results for the three different sample sizes when the number correct score was used as the conditioning variable. As was predicted, for each sample size the number correct score correlated most highly for the linear $\theta_1$, $\theta_2$ composite that represented an equal weighting of both dimensions. The correlations were .78, .73, .69 for the 1000/500, 1000/250 and 500/250 samples sizes respectively. Also, as predicted items that were measuring mostly $\theta_1$ or mostly $\theta_2$ were flagged as being biased. In the 1000/500 case, items 1-9 and 22-30 were flagged 100% of the time as significantly favoring the reference and focal groups respectively. Similar results were reported for the 1000/250 and 500/250 conditions. Both DIF statistics, the MH and $B_{uni}$ performed equally well.

Insert Tables 1-3 about here

*Hypothesis Two:*

When conditioning only on $\theta_1$ (Tables 4-6) items with angular composites greater than 30° (items 11-30) were consistently flagged by both DIF statistics as being biased against the focal group by both statistics for each sample size condition. Slight differences were noted between MH and $B_{uni}$. $B_{uni}$ appeared to be more sensitive. That is, the rejection rate for $B_{uni}$ increased at a faster rate than MH as the angular composite of the item departed from 0°.

Insert Tables 4-6 about here

*Hypothesis Three:*

As was predicted the opposite results occurred when the valid test direction was along the $\theta_2$ axis. These results are shown in Tables 7-9. When a linear transformation of $\theta_2$ was used as the matching criterion, items with angular composites less than 60° (items 1-20) were consistently flagged by MH and $B_{uni}$ as significantly favoring the reference group. As the sample size increased, the sensitivity of the statistics to depart slightly from the valid test direction increased. Again $B_{uni}$ had higher rejection rates than did the MH statistic.

Insert Tables 7-9 about here

*Hypothesis Four*

For the final set of analyses the conditioning variable identified the complete latent space. Unlike the analyses in which a single conditioning variable was used, neither the mean of the MH or the mean of $B_{uni}$ were statistically significant for any of the items. As was seen with the previous analyses $B_{uni}$ produced greater rates of Type I error for each of the sample size combinations.

Insert Tables 10-12 about here

## Discussion

Differences between the performance of MH and $B_{uni}$, especially for the final set of analyses, may be due to the way the statistics are computed. No purification process was used for the MH computations and thus the conditioning score always contained the influence of more biased items than just the studied item. In fairness to the Shealy Stout SIBTEST statistic, it needs to be noted that the required regression correction was not used. It was not necessary when the conditioning variable was a transformation of $\theta_1$, or of $\theta_{2,}$ or of $\theta_1$ and $\theta_2$, because the true underlying ability was known. However, when the coarse $(\theta_1,\theta_2)$ matching achieved by the 8 X 8 grid or when the conditioning variable was the number correct score, the higher rates of rejection (when compared to those for the MH statistic) are believed to be caused by not using the regression correction. That is, in the case of the 8 X 8 grid matching, because the size of each of the 64 two-dimensional cells were not small enough, the $(\theta_1,\theta_2)$ distributions for the

Reference and Focal groups within each cell were probably dissimilar enough to inflate the rejection rates because bias was being confounded with impact. As such, the SIBTEST analysis would have benefitted from the use of the regression correction.

In the analyses for hypotheses one, two, and three the size or amount of bias exhibited increased as the item's angular composite direction departed from the valid test direction. It also appears that for large sample sizes (i.e., 1000/500) the power of the MH and SIBTEST procedures to detect DIF increases 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°).

Determining what variables or scores to condition on to account for the complete latent ability space is obviously no easy task. However, as this study demonstrates, the process of conditioning on those variables which span the complete latent space can successfully eliminate differential performance between two groups of interest. As mentioned earlier this study represents the ideal case. In this study the underlying abilities were known. It is doubtful that practitioners will ever be able to account for, or be able to condition upon, scores which can account for the complete latent space. But by identifying scores or variables that decrease the amount of DIF, researchers can obtain a better understanding about what their test is actually measuring. This process appears to be a needed step in the direction of establishing a congruence between the content specifications from which a test is created (and is purportedly measuring) and subsequent statistical analyses which represent the actual skills employed by the examinees.

# References

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

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 and H.I. Braun (Eds.). Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Roussos, L.A., Stout, W. F. & Marden, J.I. (1993). 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. (1993). An item response theory model for test bias. In P. Holland and H. Wainer (Eds.) Differential Item Functioning (pp. 197-239) Hillsdale, NJ: Lawrence Erlbaum Associates

Shealy, R. & Stout, W. (In press). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. Psychometrika

Shin, S. (1992). An empirical investigation of the robustness of the Mantel Haenszel Procedure and sources of DIF. Unpublished doctoral dissertation, University of Illinois, Champaign, IL.

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

Table 1

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 500/250 case when the conditioning variable was the number correct score.

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 49.14 | 13.78 | .00 | 1.00 | -7.54 | 1.23 | .00 | 1.00 |
| 2 | 42.28 | 11.74 | .00 | 1.00 | -7.03 | 1.12 | .00 | 1.00 |
| 3 | 37.26 | 11.08 | .00 | 1.00 | -6.54 | 1.07 | .00 | 1.00 |
| 4 | 34.06 | 10.91 | .00 | 1.00 | -6.25 | 1.10 | .00 | 1.00 |
| 5 | 30.22 | 11.51 | .00 | 1.00 | -5.82 | 1.11 | .00 | 1.00 |
| 6 | 25.99 | 9.83 | .00 | .99 | -5.42 | 1.13 | .00 | .99 |
| 7 | 20.10 | 7.96 | .00 | 1.00 | -4.75 | 1.02 | .00 | 1.00 |
| 8 | 17.54 | 8.89 | .00 | .99 | -4.39 | 1.16 | .00 | .99 |
| 9 | 12.42 | 6.20 | .00 | .93 | -3.69 | .99 | .00 | .95 |
| 10 | 8.81 | 5.57 | .00 | .79 | -3.11 | 1.05 | .00 | .86 |
| 11 | 6.15 | 4.14 | .00 | .68 | -2.59 | 0.93 | .00 | .73 |
| 12 | 4.46 | 3.64 | .00 | .51 | -2.07 | 1.08 | .00 | .57 |
| 13 | 2.68 | 3.22 | .00 | .27 | -1.48 | 1.05 | .00 | .32 |
| 14 | 1.63 | 2.06 | .00 | .13 | -1.00 | 1.05 | .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 | .00 | .96 | 1.12 | .16 | .02 |
| 18 | 2.91 | 2.97 | .29 | .00 | 1.60 | 1.09 | .36 | .00 |
| 19 | 4.03 | 3.70 | .36 | .00 | 2.05 | 1.05 | .54 | .00 |
| 20 | 7.14 | 5.43 | .68 | .00 | 2.84 | 1.15 | .78 | .00 |
| 21 | 9.19 | 6.01 | .79 | .00 | 3.25 | 1.18 | .87 | .00 |
| 22 | 12.38 | 6.12 | .94 | .00 | 3.88 | 1.05 | .96 | .00 |
| 23 | 15.88 | 7.78 | .96 | .00 | 4.39 | 1.16 | .97 | .00 |
| 24 | 20.50 | 7.75 | 1.00 | .00 | 5.12 | 1.09 | 1.00 | .00 |
| 25 | 23.45 | 9.20 | 1.00 | .00 | 5.50 | 1.24 | 1.00 | .00 |
| 26 | 28.28 | 10.06 | 1.00 | .00 | 6.05 | 1.22 | 1.00 | .00 |
| 27 | 31.19 | 9.20 | 1.00 | .00 | 6.34 | 1.09 | 1.00 | .00 |
| 28 | 34.97 | 10.21 | 1.00 | .00 | 6.77 | 1.15 | 1.00 | .00 |
| 29 | 42.19 | 13.29 | 1.00 | .00 | 7.46 | 1.41 | 1.00 | .00 |
| 30 | 46.74 | 11.52 | 1.00 | .00 | 7.96 | 1.17 | 1.00 | .00 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 2

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/250 case when the conditioning variable was the number correct score.

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | 60.75 | 14.91 | .00 | 1.00 | -8.03 | 1.11 | .00 | 1.00 |
| 2 | 55.29 | 13.68 | .00 | 1.00 | -7.64 | .99 | .00 | 1.00 |
| 3 | 48.40 | 12.93 | .00 | 1.00 | -7.13 | 1.05 | .00 | 1.00 |
| 4 | 42.02 | 12.64 | .00 | 1.00 | -6.67 | 1.01 | .00 | 1.00 |
| 5 | 39.84 | 12.79 | .00 | 1.00 | -6.51 | 1.06 | .00 | 1.00 |
| 6 | 32.04 | 10.50 | .00 | 1.00 | -5.78 | 1.00 | .00 | 1.00 |
| 7 | 24.26 | 10.40 | .00 | .99 | -5.04 | 1.17 | .00 | .99 |
| 8 | 21.18 | 8.06 | .00 | .98 | -4.73 | 1.03 | .00 | .99 |
| 9 | 15.22 | 7.94 | .00 | .96 | -3.95 | 1.07 | .00 | .97 |
| 10 | 11.91 | 7.00 | .00 | .84 | -3.48 | 1.12 | .00 | .90 |
| 11 | 7.43 | 5.23 | .00 | .75 | -2.72 | 1.03 | .00 | .80 |
| 12 | 5.10 | 3.88 | .00 | .52 | -2.22 | .98 | .00 | .59 |
| 13 | 3.14 | 4.02 | .00 | .29 | -1.58 | 1.41 | .00 | .36 |
| 14 | 1.85 | 2.52 | .00 | .13 | -1.12 | 1.02 | .01 | .16 |
| 15 | 1.02 | 1.32 | .01 | .05 | -.30 | 1.10 | .01 | .06 |
| 16 | .90 | 1.29 | .05 | .00 | .33 | 1.05 | .08 | .00 |
| 17 | 1.47 | 1.91 | .12 | .00 | .95 | 1.02 | .17 | .00 |
| 18 | 3.41 | 3.66 | .35 | .00 | 1.77 | 1.14 | .43 | .00 |
| 19 | 5.07 | 3.63 | .56 | .00 | 2.33 | 1.07 | .64 | .00 |
| 20 | 6.68 | 4.57 | .68 | .00 | 2.75 | 1.05 | .78 | .00 |
| 21 | 10.56 | 6.38 | .85 | .00 | 3.51 | 1.15 | .90 | .00 |
| 22 | 15.43 | 7.29 | .95 | .00 | 4.31 | 1.14 | .98 | .00 |
| 23 | 19.34 | 7.94 | .98 | .00 | 4.96 | 1.16 | .99 | .00 |
| 24 | 23.15 | 8.86 | 1.00 | .00 | 5.41 | 1.20 | 1.00 | .00 |
| 25 | 28.96 | 10.30 | 1.00 | .00 | 6.10 | 1.27 | 1.00 | .00 |
| 26 | 36.14 | 10.37 | 1.00 | .00 | 6.90 | 1.18 | 1.00 | .00 |
| 27 | 38.7 | 9.94 | 1.00 | .00 | 7.16 | 1.14 | 1.00 | .00 |
| 28 | 43.73 | 11.42 | 1.00 | .00 | 7.65 | 1.21 | 1.00 | .00 |
| 29 | 50.78 | 13.19 | 1.00 | .00 | 8.27 | 1.35 | 1.00 | .00 |
| 30 | 56.46 | 13.96 | 1.00 | .00 | 8.78 | 1.37 | 1.00 | .00 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 3

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/500 case when the conditioning variable was the number correct score.

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | 100.09 | 17.51 | .00 | 1.00 | -10.46 | 1.03 | .00 | 1.00 |
| 2 | 88.25 | 18.01 | .00 | 1.00 | -9.81 | 1.03 | .00 | 1.00 |
| 3 | 79.75 | 17.94 | .00 | 1.00 | -9.28 | 1.13 | .00 | 1.00 |
| 4 | 64.89 | 13.63 | .00 | 1.00 | -8.37 | .99 | .00 | 1.00 |
| 5 | 63.06 | 14.57 | .00 | 1.00 | -8.21 | .98 | .00 | 1.00 |
| 6 | 51.78 | 12.48 | .00 | 1.00 | -7.45 | .89 | .00 | 1.00 |
| 7 | 40.41 | 11.85 | .00 | 1.00 | -6.52 | .99 | .00 | 1.00 |
| 8 | 34.41 | 10.96 | .00 | 1.00 | -6.02 | 1.01 | .00 | 1.00 |
| 9 | 26.55 | 9.75 | .00 | 1.00 | -5.28 | 1.03 | .00 | 1.00 |
| 10 | 18.91 | 9.66 | .00 | .98 | -4.39 | 1.11 | .00 | .99 |
| 11 | 11.46 | 6.36 | .00 | .90 | -3.38 | .98 | .00 | .92 |
| 12 | 8.11 | 5.03 | .00 | .81 | -2.86 | .94 | .00 | .83 |
| 13 | 4.21 | 3.26 | .00 | .50 | -1.98 | .92 | .00 | .51 |
| 14 | 1.96 | 2.01 | .00 | .20 | -1.25 | .80 | .00 | .21 |
| 15 | .99 | 1.37 | .01 | .05 | -.35 | 1.04 | .03 | .05 |
| 16 | .76 | 1.08 | .02 | .00 | .33 | .91 | .03 | .00 |
| 17 | 1.85 | 2.46 | .13 | .00 | 1.15 | .93 | .13 | .00 |
| 18 | 5.34 | 4.93 | .51 | .00 | 2.15 | 1.25 | .56 | .00 |
| 19 | 7.98 | 5.59 | .76 | .00 | 2.85 | 1.10 | .80 | .00 |
| 20 | 13.30 | 6.93 | .90 | .00 | 3.75 | 1.11 | .91 | .00 |
| 21 | 16.91 | 8.20 | .98 | .00 | 4.28 | 1.07 | .99 | .00 |
| 22 | 24.68 | 8.92 | 1.00 | .00 | 5.30 | 1.01 | 1.00 | .00 |
| 23 | 32.13 | 9.70 | 1.00 | .00 | 6.11 | 1.02 | 1.00 | .00 |
| 24 | 39.84 | 12.35 | 1.00 | .00 | 6.82 | 1.14 | 1.00 | .00 |
| 25 | 48.55 | 13.03 | 1.00 | .00 | 7.59 | 1.12 | 1.00 | .00 |
| 26 | 57.32 | 13.15 | 1.00 | .00 | 8.30 | 1.05 | 1.00 | .00 |
| 27 | 65.46 | 14.82 | 1.00 | .00 | 8.89 | 1.17 | 1.00 | .00 |
| 28 | 74.62 | 17.63 | 1.00 | .00 | 9.54 | 1.28 | 1.00 | .00 |
| 29 | 85.90 | 18.17 | 1.00 | .00 | 10.28 | 1.23 | 1.00 | .00 |
| 30 | 94.69 | 19.60 | 1.00 | .00 | 10.85 | 1.28 | 1.00 | .00 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)

[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 4

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 500/250 case when the conditioning variable was a transformation of $\theta_1$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | .90 | 1.37 | .01 | .01 | .55 | 1.56 | .21 | .03 |
| 2 | .95 | 1.32 | .07 | .00 | 1.31 | 1.35 | .30 | .01 |
| 3 | 1.54 | 2.12 | .13 | .00 | 1.87 | 1.35 | .43 | .00 |
| 4 | 2.55 | 2.62 | .21 | .00 | 2.58 | 1.29 | .68 | .00 |
| 5 | 3.68 | 3.55 | .39 | .00 | 2.97 | 1.50 | .77 | .00 |
| 6 | 5.02 | 4.42 | .50 | .00 | 3.49 | 1.75 | .82 | .00 |
| 7 | 7.65 | 4.90 | .74 | .00 | 4.38 | 1.55 | .93 | .00 |
| 8 | 9.03 | 5.93 | .86 | .00 | 4.70 | 1.77 | .95 | .00 |
| 9 | 11.67 | 6.11 | .93 | .00 | 5.30 | 1.70 | .97 | .00 |
| 10 | 14.89 | 6.34 | .97 | .00 | 5.99 | 1.36 | 1.00 | .00 |
| 11 | 16.76 | 5.89 | .99 | .00 | 6.40 | 1.59 | 1.00 | .00 |
| 12 | 18.89 | 7.17 | 1.00 | .00 | 6.66 | 1.62 | .99 | .00 |
| 13 | 22.17 | 7.85 | 1.00 | .00 | 7.36 | 1.61 | 1.00 | .00 |
| 14 | 23.72 | 7.74 | 1.00 | .00 | 7.58 | 1.65 | 1.00 | .00 |
| 15 | 26.15 | 8.82 | 1.00 | .00 | 8.09 | 1.79 | 1.00 | .00 |
| 16 | 28.32 | 8.67 | 1.00 | .00 | 8.27 | 1.57 | 1.00 | .00 |
| 17 | 31.15 | 10.01 | 1.00 | .00 | 8.60 | 1.88 | 1.00 | .00 |
| 18 | 35.32 | 10.98 | 1.00 | .00 | 9.26 | 2.01 | 1.00 | .00 |
| 19 | 35.23 | 9.79 | 1.00 | .00 | 9.27 | 1.70 | 1.00 | .00 |
| 20 | 38.41 | 10.51 | 1.00 | .00 | 9.63 | 1.85 | 1.00 | .00 |
| 21 | 40.76 | 11.34 | 1.00 | .00 | 9.94 | 1.83 | 1.00 | .00 |
| 22 | 42.88 | 10.37 | 1.00 | .00 | 10.02 | 1.64 | 1.00 | .00 |
| 23 | 43.76 | 11.78 | 1.00 | .00 | 10.13 | 1.87 | 1.00 | .00 |
| 24 | 45.78 | 11.18 | 1.00 | .00 | 10.26 | 1.74 | 1.00 | .00 |
| 25 | 47.61 | 11.16 | 1.00 | .00 | 10.48 | 1.82 | 1.00 | .00 |
| 26 | 48.45 | 11.38 | 1.00 | .00 | 10.31 | 1.67 | 1.00 | .00 |
| 27 | 48.63 | 11.81 | 1.00 | .00 | 10.34 | 1.70 | 1.00 | .00 |
| 28 | 48.26 | 11.28 | 1.00 | .00 | 10.22 | 1.74 | 1.00 | .00 |
| 29 | 51.32 | 13.49 | 1.00 | .00 | 10.57 | 1.82 | 1.00 | .00 |
| 30 | 53.17 | 13.13 | 1.00 | .00 | 10.54 | 1.82 | 1.00 | .00 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 5

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/250 case when the conditioning variable was a transformation of $\theta_1$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | .76 | 1.24 | .01 | .01 | .48 | 1.32 | .13 | .01 |
| 2 | 1.09 | 1.70 | .05 | .01 | 1.23 | 1.42 | .25 | .02 |
| 3 | 1.87 | 2.33 | .16 | .00 | 1.92 | 1.36 | .44 | .00 |
| 4 | 3.21 | 2.89 | .33 | .00 | 2.69 | 1.45 | .68 | .00 |
| 5 | 4.27 | 3.69 | .42 | .00 | 3.04 | 1.48 | .75 | .00 |
| 6 | 6.53 | 5.21 | .65 | .00 | 3.79 | 1.67 | .87 | .00 |
| 7 | 10.87 | 5.92 | .91 | .00 | 5.01 | 1.65 | .97 | .00 |
| 8 | 11.45 | 6.04 | .93 | .00 | 5.21 | 1.60 | .98 | .00 |
| 9 | 15.47 | 7.93 | .97 | .00 | 5.98 | 1.74 | .99 | .00 |
| 10 | 18.96 | 8.99 | .98 | .00 | 6.73 | 1.85 | 1.00 | .00 |
| 11 | 23.17 | 8.37 | 1.00 | .00 | 7.48 | 1.76 | 1.00 | .00 |
| 12 | 24.67 | 8.63 | 1.00 | .00 | 7.72 | 1.89 | 1.00 | .00 |
| 13 | 28.82 | 10.00 | 1.00 | .00 | 8.47 | 2.03 | 1.00 | .00 |
| 14 | 30.20 | 9.59 | 1.00 | .00 | 8.64 | 1.85 | 1.00 | .00 |
| 15 | 34.34 | 10.07 | 1.00 | .00 | 9.28 | 1.89 | 1.00 | .00 |
| 16 | 36.66 | 10.66 | 1.00 | .00 | 9.46 | 1.74 | 1.00 | .00 |
| 17 | 39.73 | 10.06 | 1.00 | .00 | 10.00 | 1.79 | 1.00 | .00 |
| 18 | 44.94 | 12.46 | 1.00 | .00 | 10.60 | 2.18 | 1.00 | .00 |
| 19 | 46.25 | 10.50 | 1.00 | .00 | 10.66 | 1.84 | 1.00 | .00 |
| 20 | 47.05 | 10.43 | 1.00 | .00 | 10.69 | 1.73 | 1.00 | .00 |
| 21 | 50.98 | 10.90 | 1.00 | .00 | 11.33 | 1.92 | 1.00 | .00 |
| 22 | 54.81 | 13.35 | 1.00 | .00 | 11.53 | 2.01 | 1.00 | .00 |
| 23 | 56.46 | 12.96 | 1.00 | .00 | 11.64 | 2.08 | 1.00 | .00 |
| 24 | 57.46 | 13.22 | 1.00 | .00 | 11.61 | 2.00 | 1.00 | .00 |
| 25 | 59.85 | 13.36 | 1.00 | .00 | 11.92 | 2.05 | 1.00 | .00 |
| 26 | 63.76 | 12.45 | 1.00 | .00 | 11.98 | 1.84 | 1.00 | .00 |
| 27 | 63.76 | 14.57 | 1.00 | .00 | 12.08 | 2.00 | 1.00 | .00 |
| 28 | 62.80 | 12.68 | 1.00 | .00 | 11.73 | 1.85 | 1.00 | .00 |
| 29 | 65.08 | 13.42 | 1.00 | .00 | 11.77 | 1.76 | 1.00 | .00 |
| 30 | 66.38 | 12.79 | 1.00 | .00 | 11.72 | 1.74 | 1.00 | .00 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)

[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 6

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/500 case when the conditioning variable was a transformation of $\theta_1$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | .79 | 1.46 | .02 | .02 | .48 | 1.27 | .13 | .03 |
| 2 | 1.37 | 1.86 | .10 | .00 | 1.52 | 1.34 | .35 | .01 |
| 3 | 2.56 | 2.69 | .29 | .00 | 2.11 | 1.49 | .55 | .00 |
| 4 | 5.59 | 4.11 | .57 | .00 | 3.35 | 1.24 | .85 | .00 |
| 5 | 6.81 | 4.55 | .70 | .00 | 3.59 | 1.46 | .83 | .00 |
| 6 | 9.80 | 5.58 | .89 | .00 | 4.31 | 1.35 | .97 | .00 |
| 7 | 15.50 | 6.86 | .99 | .00 | 5.49 | 1.53 | 1.00 | .00 |
| 8 | 18.22 | 8.22 | .99 | .00 | 6.00 | 1.63 | .99 | .00 |
| 9 | 22.57 | 8.72 | 1.00 | .00 | 6.60 | 1.46 | 1.00 | .00 |
| 10 | 29.27 | 10.88 | 1.00 | .00 | 7.61 | 1.57 | 1.00 | .00 |
| 11 | 36.35 | 10.32 | 1.00 | .00 | 8.59 | 1.46 | 1.00 | .00 |
| 12 | 39.41 | 11.46 | 1.00 | .00 | 8.78 | 1.75 | 1.00 | .00 |
| 13 | 44.27 | 10.51 | 1.00 | .00 | 9.53 | 1.27 | 1.00 | .00 |
| 14 | 47.13 | 10.97 | 1.00 | .00 | 9.74 | 1.44 | 1.00 | .00 |
| 15 | 54.46 | 13.39 | 1.00 | .00 | 10.53 | 1.69 | 1.00 | .00 |
| 16 | 57.61 | 13.13 | 1.00 | .00 | 10.72 | 1.61 | 1.00 | .00 |
| 17 | 62.58 | 12.68 | 1.00 | .00 | 11.14 | 1.51 | 1.00 | .00 |
| 18 | 70.00 | 17.23 | 1.00 | .00 | 11.86 | 2.00 | 1.00 | .00 |
| 19 | 72.27 | 14.69 | 1.00 | .00 | 12.11 | 1.75 | 1.00 | .00 |
| 20 | 78.79 | 17.08 | 1.00 | .00 | 12.50 | 1.84 | 1.00 | .00 |
| 21 | 79.05 | 14.61 | 1.00 | .00 | 12.60 | 1.55 | 1.00 | .00 |
| 22 | 85.48 | 16.19 | 1.00 | .00 | 12.97 | 1.74 | 1.00 | .00 |
| 23 | 89.39 | 15.80 | 1.00 | .00 | 13.16 | 1.71 | 1.00 | .00 |
| 24 | 90.73 | 17.65 | 1.00 | .00 | 13.08 | 1.91 | 1.00 | .00 |
| 25 | 95.33 | 16.55 | 1.00 | .00 | 13.29 | 1.71 | 1.00 | .00 |
| 26 | 98.32 | 16.15 | 1.00 | .00 | 13.45 | 1.64 | 1.00 | .00 |
| 27 | 99.02 | 15.93 | 1.00 | .00 | 13.27 | 1.49 | 1.00 | .00 |
| 28 | 100.77 | 18.65 | 1.00 | .00 | 13.26 | 1.86 | 1.00 | .00 |
| 29 | 105.49 | 18.84 | 1.00 | .00 | 13.48 | 1.60 | 1.00 | .00 |
| 30 | 106.03 | 18.58 | 1.00 | .00 | 13.29 | 1.56 | 1.00 | .00 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 7

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 500/250 case when the conditioning variable was a transformation of $\theta_2$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | 58.11 | 15.00 | .00 | 1.00 | -9.03 | 1.45 | .00 | 1.00 |
| 2 | 57.52 | 14.36 | .00 | 1.00 | -8.93 | 1.38 | .00 | 1.00 |
| 3 | 56.66 | 14.77 | .00 | 1.00 | -8.91 | 1.46 | .00 | 1.00 |
| 4 | 56.45 | 13.54 | .00 | 1.00 | -8.81 | 1.31 | .00 | 1.00 |
| 5 | 54.80 | 14.86 | .00 | 1.00 | -8.57 | 1.45 | .00 | 1.00 |
| 6 | 54.37 | 13.56 | .00 | 1.00 | -8.50 | 1.23 | .00 | 1.00 |
| 7 | 50.19 | 12.79 | .00 | 1.00 | -8.14 | 1.20 | .00 | 1.00 |
| 8 | 50.84 | 13.90 | .00 | 1.00 | -8.08 | 1.35 | .00 | 1.00 |
| 9 | 46.75 | 12.73 | .00 | 1.00 | -7.68 | 1.33 | .00 | 1.00 |
| 10 | 44.99 | 12.25 | .00 | 1.00 | -7.43 | 1.21 | .00 | 1.00 |
| 11 | 42.31 | 10.25 | .00 | 1.00 | -7.19 | 1.14 | .00 | 1.00 |
| 12 | 41.11 | 12.09 | .00 | 1.00 | -6.97 | 1.28 | .00 | 1.00 |
| 13 | 37.19 | 12.23 | .00 | 1.00 | -6.62 | 1.30 | .00 | 1.00 |
| 14 | 35.21 | 11.66 | .00 | 1.00 | -6.42 | 1.29 | .00 | 1.00 |
| 15 | 32.07 | 10.21 | .00 | 1.00 | -6.02 | 1.16 | .00 | 1.00 |
| 16 | 29.01 | 10.12 | .00 | 1.00 | -5.68 | 1.21 | .00 | 1.00 |
| 17 | 25.66 | 10.87 | .00 | 1.00 | -5.15 | 1.30 | .00 | .99 |
| 18 | 22.96 | 8.68 | .00 | 1.00 | -4.91 | 1.02 | .00 | 1.00 |
| 19 | 21.83 | 9.21 | .00 | .98 | -4.84 | 1.12 | .00 | .99 |
| 20 | 17.47 | 8.36 | .00 | .99 | -4.20 | 1.10 | .00 | .99 |
| 21 | 14.70 | 7.52 | .00 | .96 | -3.85 | 1.13 | .00 | .97 |
| 22 | 11.26 | 5.63 | .00 | .93 | -3.39 | .94 | .00 | .96 |
| 23 | 9.90 | 6.12 | .00 | .86 | -3.06 | 1.06 | .00 | .87 |
| 24 | 7.26 | 4.82 | .00 | .69 | -2.60 | 1.05 | .00 | .72 |
| 25 | 5.93 | 4.51 | .00 | .57 | -2.32 | 1.07 | .00 | .57 |
| 26 | 3.73 | 3.37 | .00 | .40 | -1.81 | 1.10 | .00 | .39 |
| 27 | 2.88 | 3.02 | .00 | .31 | -1.52 | 1.09 | .00 | .39 |
| 28 | 1.77 | 2.15 | .00 | .13 | -1.14 | .89 | .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 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)

[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 8

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/250 case when the conditioning variable was a transformation of $\theta_2$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | 78.47 | 18.73 | .00 | 1.00 | -9.90 | 1.41 | .00 | 1.00 |
| 2 | 79.02 | 17.57 | .00 | 1.00 | -9.91 | 1.46 | .00 | 1.00 |
| 3 | 77.66 | 17.53 | .00 | 1.00 | -9.65 | 1.39 | .00 | 1.00 |
| 4 | 75.30 | 16.35 | .00 | 1.00 | -9.48 | 1.31 | .00 | 1.00 |
| 5 | 78.21 | 19.62 | .00 | 1.00 | -9.67 | 1.55 | .00 | 1.00 |
| 6 | 73.97 | 16.79 | .00 | 1.00 | -9.18 | 1.35 | .00 | 1.00 |
| 7 | 68.11 | 18.03 | .00 | 1.00 | -8.73 | 1.55 | .00 | 1.00 |
| 8 | 68.20 | 15.46 | .00 | 1.00 | -8.63 | 1.26 | .00 | 1.00 |
| 9 | 62.82 | 16.18 | .00 | 1.00 | -8.18 | 1.25 | .00 | 1.00 |
| 10 | 61.87 | 16.56 | .00 | 1.00 | -7.93 | 1.30 | .00 | 1.00 |
| 11 | 55.19 | 13.67 | .00 | 1.00 | -7.53 | 1.10 | .00 | 1.00 |
| 12 | 52.95 | 14.56 | .00 | 1.00 | -7.26 | 1.34 | .00 | 1.00 |
| 13 | 49.68 | 14.28 | .00 | 1.00 | -7.07 | 1.25 | .00 | 1.00 |
| 14 | 47.75 | 14.40 | .00 | 1.00 | -6.83 | 1.34 | .00 | 1.00 |
| 15 | 42.56 | 13.27 | .00 | 1.00 | -6.41 | 1.16 | .00 | 1.00 |
| 16 | 38.81 | 13.61 | .00 | 1.00 | -6.10 | 1.36 | .00 | 1.00 |
| 17 | 34.67 | 12.00 | .00 | 1.00 | -5.69 | 1.12 | .00 | 1.00 |
| 18 | 30.20 | 11.94 | .00 | 1.00 | -5.25 | 1.18 | .00 | 1.00 |
| 19 | 26.51 | 10.32 | .00 | 1.00 | -4.93 | 1.11 | .00 | 1.00 |
| 20 | 24.53 | 8.80 | .00 | 1.00 | -4.72 | 1.00 | .00 | .99 |
| 21 | 19.93 | 8.80 | .00 | .98 | -4.21 | 1.10 | .00 | .98 |
| 22 | 14.47 | 7.58 | .00 | .95 | -3.58 | 1.09 | .00 | .94 |
| 23 | 12.27 | 6.63 | .00 | .93 | -3.26 | .99 | .00 | .94 |
| 24 | 10.03 | 6.94 | .00 | .80 | -2.98 | 1.09 | .00 | .79 |
| 25 | 6.90 | 5.54 | .00 | .64 | -2.36 | 1.12 | .00 | .64 |
| 26 | 4.14 | 3.50 | .00 | .46 | -1.81 | .97 | .00 | .46 |
| 27 | 3.07 | 3.26 | .00 | .32 | -1.50 | .96 | .00 | .33 |
| 28 | 1.86 | 1.90 | .00 | .17 | -1.11 | .87 | .00 | .22 |
| 29 | 1.18 | 1.37 | .02 | .06 | -.46 | .02 | .02 | .08 |
| 30 | .70 | 1.02 | .01 | .01 | -.01 | .95 | .01 | .03 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 9

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/500 case when the conditioning variable was a transformation of $\theta_2$

| Item | $\bar{X}_{MH}$ | $_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | 124.58 | 19.55 | .00 | 1.00 | -12.12 | 1.31 | .00 | 1.00 |
| 2 | 121.52 | 20.36 | .00 | 1.00 | -12.00 | 1.28 | .00 | 1.00 |
| 3 | 121.85 | 20.74 | .00 | 1.00 | -11.92 | 1.29 | .00 | 1.00 |
| 4 | 112.24 | 18.13 | .00 | 1.00 | -11.41 | 1.10 | .00 | 1.00 |
| 5 | 116.87 | 19.58 | .00 | 1.00 | -11.44 | 1.20 | .00 | 1.00 |
| 6 | 112.90 | 18.81 | .00 | 1.00 | -11.30 | 1.31 | .00 | 1.00 |
| 7 | 105.80 | 17.01 | .00 | 1.00 | -10.72 | 1.21 | .00 | 1.00 |
| 8 | 104.03 | 18.68 | .00 | 1.00 | -10.60 | 1.17 | .00 | 1.00 |
| 9 | 99.83 | 17.23 | .00 | 1.00 | -10.31 | 1.14 | .00 | 1.00 |
| 10 | 92.90 | 18.56 | .00 | 1.00 | -9.79 | 1.19 | .00 | 1.00 |
| 11 | 84.79 | 19.10 | .00 | 1.00 | -9.30 | 1.26 | .00 | 1.00 |
| 12 | 84.36 | 14.98 | .00 | 1.00 | -9.16 | 1.05 | .00 | 1.00 |
| 13 | 77.90 | 15.40 | .00 | 1.00 | -8.75 | 1.01 | .00 | 1.00 |
| 14 | 73.61 | 15.22 | .00 | 1.00 | -8.45 | 1.05 | .00 | 1.00 |
| 15 | 65.70 | 12.66 | .00 | 1.00 | -7.92 | 1.02 | .00 | 1.00 |
| 16 | 60.79 | 14.05 | .00 | 1.00 | -7.60 | .99 | .00 | 1.00 |
| 17 | 55.49 | 12.53 | .00 | 1.00 | -7.16 | .99 | .00 | 1.00 |
| 18 | 47.50 | 15.19 | .00 | 1.00 | -6.57 | 1.14 | .00 | 1.00 |
| 19 | 42.86 | 13.24 | .00 | 1.00 | -6.17 | 1.07 | .00 | 1.00 |
| 20 | 36.09 | 12.71 | .00 | 1.00 | -5.63 | 1.05 | .00 | 1.00 |
| 21 | 32.08 | 10.66 | .00 | .99 | -5.31 | 1.06 | .00 | .99 |
| 22 | 25.01 | 8.81 | .00 | 1.00 | -4.68 | .96 | .00 | 1.00 |
| 23 | 20.05 | 8.55 | .00 | 1.00 | -4.12 | .98 | .00 | .99 |
| 24 | 14.74 | 7.30 | .00 | .98 | -3.48 | 1.05 | .00 | .91 |
| 25 | 11.05 | 5.90 | .00 | .93 | -3.05 | .95 | .00 | .87 |
| 26 | 7.43 | 4.57 | .00 | .73 | -2.45 | .89 | .00 | .71 |
| 27 | 4.86 | 4.01 | .00 | .54 | -1.85 | 1.06 | .00 | .53 |
| 28 | 2.96 | 3.79 | .00 | .28 | -1.28 | 1.08 | .00 | .26 |
| 29 | 1.32 | 1.65 | .01 | .10 | -.54 | .98 | .01 | .08 |
| 30 | .91 | 1.28 | .04 | .01 | .04 | 1.02 | .04 | .02 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 10

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 500/250 case when the conditioning variable was $\theta_1$ and $\theta_2$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | .95 | 1.28 | .00 | .06 | -.24 | 2.24 | .18 | .22 |
| 2 | .74 | 1.03 | .00 | .02 | -.44 | 2.26 | .15 | .28 |
| 3 | .72 | 1.05 | .01 | .01 | -.32 | 1.91 | .13 | .22 |
| 4 | .64 | .91 | .00 | .03 | -.56 | 1.83 | .13 | .23 |
| 5 | .91 | 1.34 | .01 | .05 | -.26 | 1.99 | .11 | .19 |
| 6 | .98 | 1.71 | .00 | .05 | -.40 | 1.87 | .12 | .18 |
| 7 | .61 | .93 | .00 | .01 | -.22 | 1.89 | .10 | .14 |
| 8 | .97 | 1.47 | .00 | .07 | -.35 | 2.20 | .18 | .26 |
| 9 | .83 | 1.13 | .00 | .02 | -.23 | 1.99 | .14 | .21 |
| 10 | .70 | 1.12 | .02 | .02 | -.12 | 2.02 | .14 | .14 |
| 11 | .63 | 1.20 | .00 | .01 | -.22 | 2.08 | .17 | .20 |
| 12 | .73 | .98 | .00 | .01 | -.42 | 1.91 | .08 | .24 |
| 13 | .66 | 1.10 | .02 | .00 | -.20 | 1.91 | .15 | .20 |
| 14 | .84 | 1.39 | .03 | .01 | -.17 | 2.15 | .16 | .20 |
| 15 | .70 | 1.00 | .01 | .00 | -.12 | 2.06 | .14 | .15 |
| 16 | .80 | 1.41 | .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 | .00 | -.24 | 1.77 | .11 | .13 |
| 20 | .68 | .92 | .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 | .00 | .19 | 1.88 | .17 | .13 |
| 28 | .81 | 1.20 | .02 | .00 | .01 | 1.80 | .17 | .14 |
| 29 | 1.19 | 1.45 | .05 | .00 | .46 | 2.03 | .21 | .09 |
| 30 | .99 | 1.51 | .04 | .01 | .37 | 1.79 | .22 | .08 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)

[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 11

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/250 case when the conditioning variable was $\theta_1$ and $\theta_2$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}^{a}$ | $P_{Foc}^{b}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.06 | 1.84 | .01 | .05 | -.66 | 2.04 | .07 | .28 |
| 2 | 1.00 | 1.66 | .01 | .05 | -.68 | 1.65 | .05 | .22 |
| 3 | .93 | 1.28 | .00 | .05 | -.66 | 1.81 | .11 | .25 |
| 4 | .79 | 1.24 | .00 | .03 | -.29 | 1.90 | .10 | .19 |
| 5 | .97 | 1.69 | .00 | .06 | -.65 | 1.87 | .07 | .25 |
| 6 | 1.20 | 1.50 | .01 | .06 | -.75 | 1.97 | .10 | .27 |
| 7 | .85 | 1.71 | .02 | .03 | -.29 | 1.66 | .10 | .15 |
| 8 | 1.09 | 1.69 | .03 | .07 | -.71 | 1.90 | .08 | .25 |
| 9 | .83 | 1.36 | .03 | .01 | -.56 | 1.70 | .10 | .24 |
| 10 | 1.01 | 1.48 | .02 | .03 | -.55 | 1.97 | .06 | .23 |
| 11 | .76 | 1.29 | .04 | .00 | -.19 | 1.81 | .11 | .19 |
| 12 | .71 | .99 | .01 | .00 | -.36 | 1.98 | .11 | .23 |
| 13 | .99 | 1.33 | .01 | .04 | -.44 | 1.82 | .10 | .18 |
| 14 | .84 | 1.13 | .01 | .01 | -.40 | 1.87 | .09 | .18 |
| 15 | .85 | 1.27 | .02 | .03 | -.42 | 1.80 | .07 | .17 |
| 16 | .92 | 1.70 | .01 | .04 | -.55 | 1.77 | .07 | .23 |
| 17 | .95 | 1.25 | .02 | .02 | -.23 | 1.95 | .15 | .21 |
| 18 | .89 | 1.51 | .03 | .01 | -.05 | 1.83 | .12 | .13 |
| 19 | .75 | 1.12 | .02 | .02 | -.25 | 1.77 | .12 | .14 |
| 20 | .62 | .89 | .01 | .00 | -.36 | 1.43 | .06 | .15 |
| 21 | .71 | 1.32 | .03 | .00 | -.31 | 1.55 | .08 | .14 |
| 22 | 1.08 | 1.65 | .04 | .02 | .11 | 1.84 | .16 | .13 |
| 23 | .78 | 1.44 | .05 | .00 | .05 | 1.64 | .13 | .09 |
| 24 | .91 | 1.42 | .05 | .03 | -.15 | 1.51 | .08 | .11 |
| 25 | .82 | 1.43 | .05 | .00 | -.14 | 1.82 | .12 | .16 |
| 26 | .82 | 1.44 | .04 | .01 | -.04 | 1.57 | .08 | .10 |
| 27 | .84 | 1.29 | .03 | .01 | -.06 | 1.36 | .05 | .08 |
| 28 | .67 | 1.08 | .01 | .00 | -.08 | 1.45 | .10 | .09 |
| 29 | .88 | 1.37 | .04 | .00 | .22 | 1.55 | .16 | .06 |
| 30 | .69 | 1.30 | .01 | .01 | .02 | 1.38 | .04 | .08 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

Table 12

Means, standard deviations, and rejection rates for the MH and $B_{uni}$ statistics for the 1000/500 case when the conditioning variable was $\theta_1$ and $\theta_2$

| Item | $\bar{X}_{MH}$ | $s_{MH}$ | $P_{Ref}$[a] | $P_{Foc}$[b] | $\bar{X}_{B_{uni}}$ | $s_{B_{uni}}$ | $P_{Ref}$[a] | $P_{Foc}$[b] |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.15 | 1.79 | .01 | .05 | -.45 | 1.78 | .12 | .21 |
| 2 | 1.14 | 1.66 | .00 | .09 | -.42 | 1.91 | .12 | .19 |
| 3 | .99 | 1.29 | .00 | .05 | -.28 | 1.89 | .12 | .16 |
| 4 | .64 | 1.07 | .02 | .00 | .20 | 1.95 | .15 | .09 |
| 5 | 1.01 | 1.31 | .01 | .03 | -.34 | 1.97 | .10 | .18 |
| 6 | 1.14 | 1.58 | .01 | .06 | -.67 | 1.75 | .08 | .26 |
| 7 | .85 | 1.22 | .01 | .02 | -.04 | 1.96 | .13 | .18 |
| 8 | 1.06 | 1.57 | .02 | .05 | -.55 | 1.97 | .15 | .27 |
| 9 | .95 | 1.21 | .00 | .03 | -.25 | 1.77 | .07 | .18 |
| 10 | 1.03 | 1.47 | .03 | .03 | -.18 | 1.81 | .11 | .17 |
| 11 | .99 | 1.13 | .02 | .01 | -.00 | 2.01 | .15 | .14 |
| 12 | .74 | .90 | .01 | .00 | -.17 | 1.85 | .16 | .18 |
| 13 | .76 | 1.15 | .01 | .03 | -.15 | 1.93 | .17 | .19 |
| 14 | .87 | 1.55 | .01 | .05 | -.16 | 1.57 | .11 | .11 |
| 15 | .78 | 1,07 | .01 | .01 | -.03 | 1.55 | .14 | .12 |
| 16 | .81 | 1.41 | .01 | .03 | -.12 | 1.68 | .10 | .13 |
| 17 | .72 | 1.48 | .02 | .02 | .10 | 1.94 | .12 | .11 |
| 18 | 1.18 | 1.62 | .02 | .04 | -.09 | 2.13 | .12 | .17 |
| 19 | .91 | 1.10 | .02 | .02 | .08 | 1.58 | .13 | .07 |
| 20 | .81 | 1.30 | .04 | .00 | .31 | 1.64 | .14 | .09 |
| 21 | .86 | 1.47 | .02 | .04 | .11 | 1.86 | .17 | .09 |
| 22 | .81 | 1.15 | .04 | .01 | .24 | 1.60 | .13 | .08 |
| 23 | .91 | 1.25 | .02 | .03 | .17 | 1.52 | .11 | .07 |
| 24 | .93 | 1.39 | .04 | .00 | .19 | 1.74 | .17 | .10 |
| 25 | .91 | 1.39 | .03 | .00 | .34 | 1.64 | .17 | .04 |
| 26 | .76 | 1.20 | .05 | .00 | .29 | 1.54 | .15 | .05 |
| 27 | 1.00 | 1.47 | .05 | .00 | .28 | 1.46 | .15 | .03 |
| 28 | .99 | 1.44 | .04 | .02 | .38 | 1.56 | .16 | .07 |
| 29 | 1.09 | 1.67 | .07 | .00 | .50 | 1.46 | .16 | .06 |
| 30 | 1.00 | 1.62 | .07 | .00 | .54 | 1.53 | .20 | .04 |

[a] Denotes percent of times the item was flagged as biased against the Reference group ($\alpha = .05$)
[b] Denotes percent of times the item was flagged as biased against the Focal group ($\alpha = .05$)

# Figure Captions

Figure 1. An item vector plot illustrating the angular composites of the 30-item simulated test.

Figure 2. Contours and marginals of the generating two-dimensional ability distributions for the reference and focal groups.

## *Figure 1*

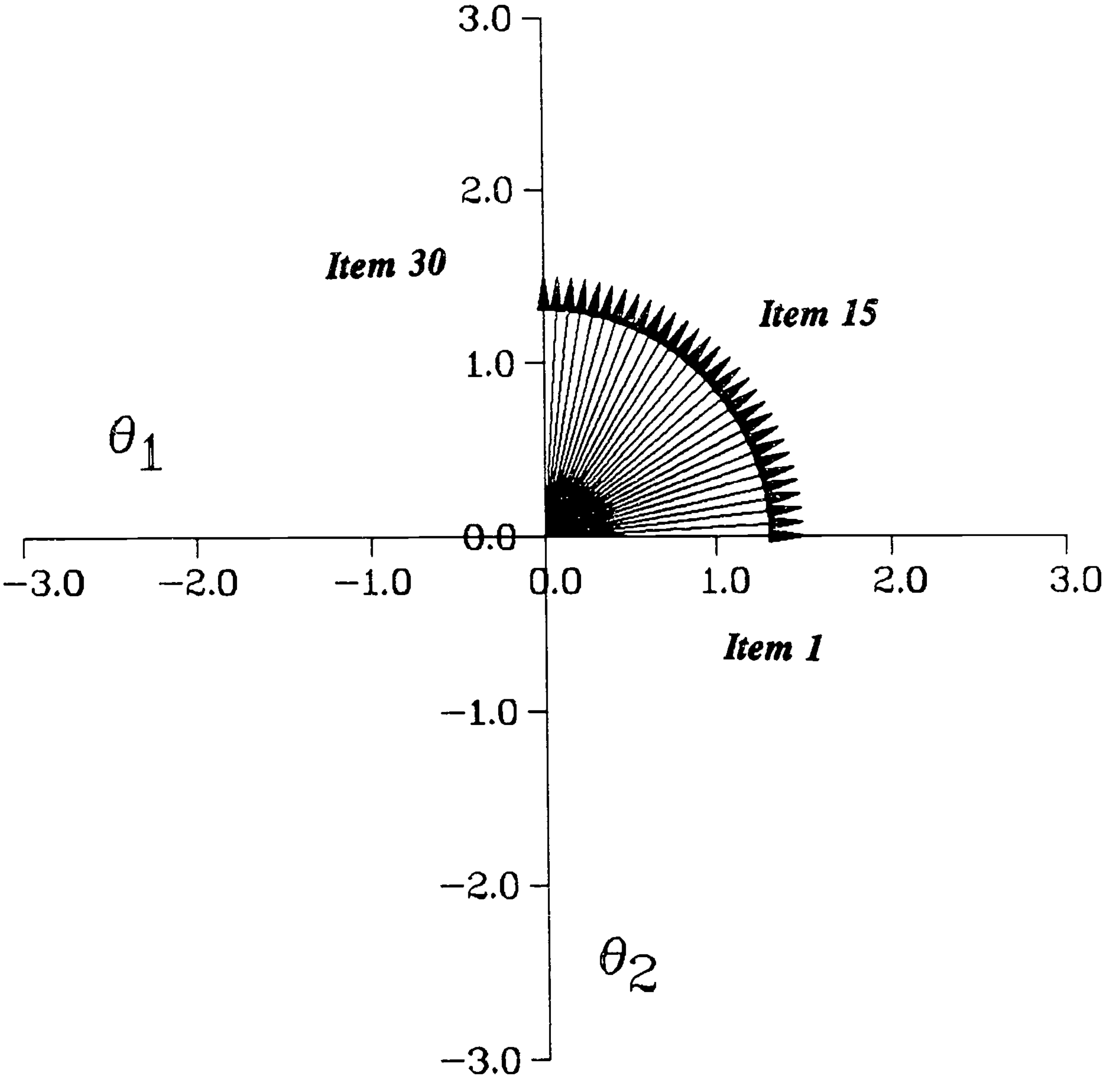
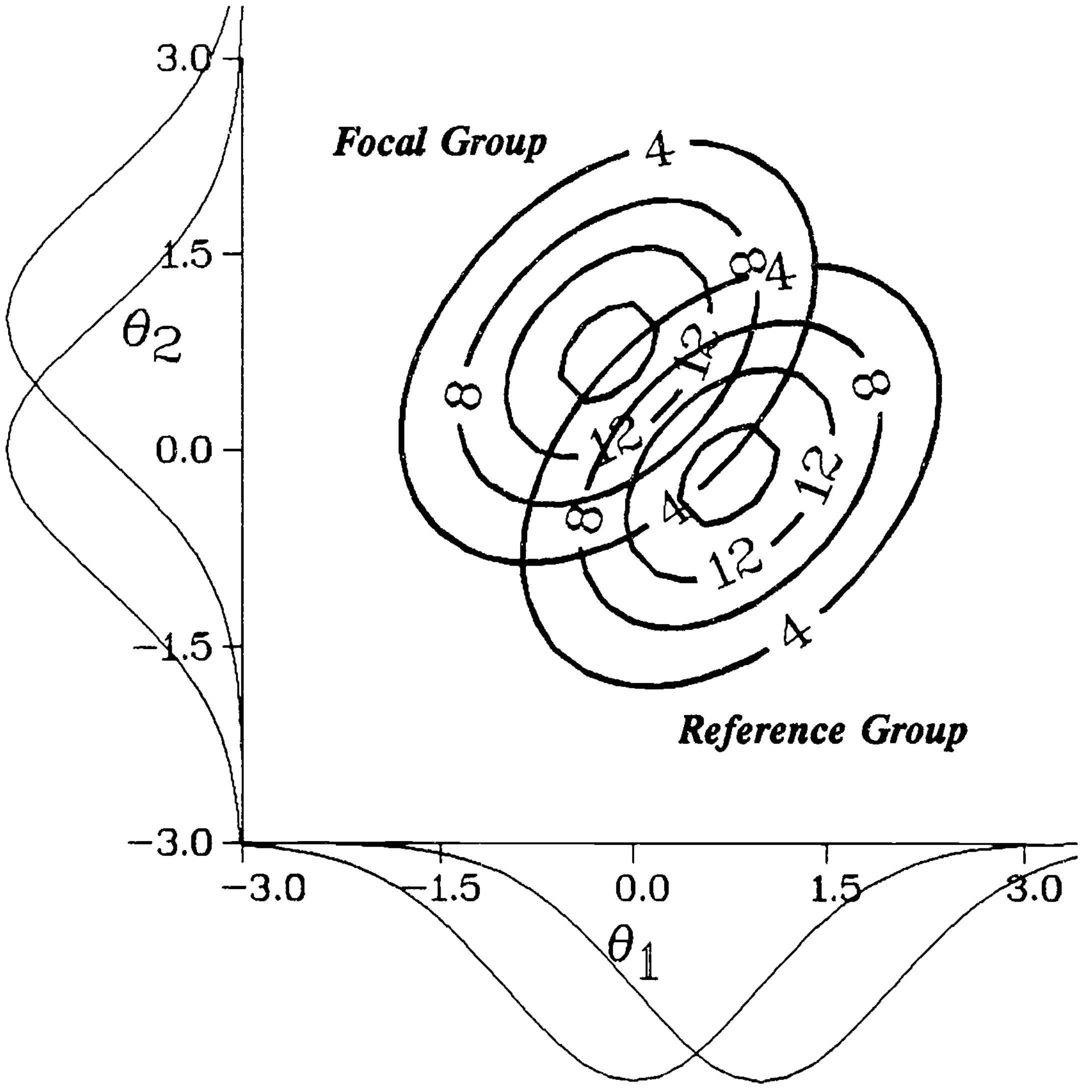

3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
Item 30
Item 15
Item 1
$\theta_1$
$\theta_2$

*Figure 2*


3.0
1.5
0.0
−1.5
−3.0
$\theta_2$
Focal Group
4
8
12
Reference Group
−3.0
−1.5
0.0
1.5
3.0
$\theta_1$