DOCUMENT RESUME

ED 298 135 TM 011 822

AUTHOR	Hambleton, Ronald K.; And Others
TITLE	Identifying Potentially Biased Test Items: A Comparison of the Mantel-Haenszel Statistic and Several Item Response Theory Methods.
SPONS AGENCY	Air Force Human Resources Lab., Brooks AFB, Texas.
PUB DATE	16 Mar 88
CONTRACT	F33615-84-C-0058
NOTE	37p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 16-20, 1986).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Cutting Scores; Grade 9; *Item Analysis; Junior High Schools; Junior High School Students; *Latent Trait Theory; Minimum Competency Testing; Reading Tests; *Sex Bias; *Test Items
IDENTIFIERS	*Mantel Haenszel Procedure; Plot Method; Route Mean Squared Difference Method; Total Area Method

ABSTRACT

Four item bias methods were studied. The methods compared include the Mantel-Haenszel statistic, the plot method, the route mean squared difference method, and the total area method; the latter two methods are based on item response theory. The test consisted of item responses of 451 male and 486 female ninth graders to 75 test items on the 1985 Cleveland Reading Competency Test. Focus was on sex bias. Simulated data were used to set cut-off scores for interpreting the item bias statistics. Each method led to the identification of nearly the same set of potentially biased items. Methodological problems included imprecision in establishing cut-off points, Type I errors, and poor item parameter estimates. Results highlight the importance of the choice of interval on the ability scale over which item bias is measured. It appears that the Mantel-Haenszel statistic provides a quick, cheap alternative to the more laborious and expensive item response theory methods. Seven graphs and four tables are included. (TJH)

Identifying Potentially Biased Test Items: A Comparison of the Mantel-Haenszel Statistic and Several Item Response Theory Methods

Ronald K. Hambleton, H. Jane Rogers, Dean Arrasmith
University of Massachusetts at Amherst

Abstract

The purpose of the present study was to compare four item bias methods, two of which are relatively new and appear to have promise, the Mantel-Haenszel (MH) statistic and the plot method, with two other promising item response theory methods, the root mean squared difference method and the total area method.

The test data consisted of the item responses of 937 ninth grade students to the 92 test items on a 1985 reading competency test. Sex bias was of principal concern in the study. Simulated data were used to set cut-off scores for interpreting the item bias statistics.

The evidence seemed clear that the four methods led, methodological problems aside, to the identification of nearly the same set of potentially biased items. Methodological problems included imprecision in establishing cut-off points, type I errors, and poor item parameter estimates. Results from the study also highlighted the significance of the choice of interval on the ability scale over which item bias is measured.

The tentative conclusion is that the Mantel-Haenszel statistic appears to provide a quick, cheap alternative to the more laborious and expensive IRT-based item bias methods. The MH results from this investigation were very supportive of the conclusion though clearly more comparisons with new datasets must be carried out before stronger conclusions are justified.

AERA in 86, Bias.1

3/16/88

Identifying Potentially Biased Test Items: A Comparison of the Mantel-Haenszel Statistic and Several Item Response Theory Methods[1,2]

Ronald K. Hambleton, H. Jane Rogers, Dean Arrasmith
University of Massachusetts at Amherst

Questions about unfairness or bias in testing have led to substantial numbers of research studies that have described and evaluated new methods for identifying potentially biased test items (see, for example, Berk, 1982; Ironson, 1982; Shepard, 1981). The purpose of the present study was to compare four different methods, two of which are relatively new and promising, the Mantel-Haenszel statistic (Holland & Thayer, 1986) and the weighted b-value plot method (Hambleton & Rogers, 1986), with two other promising item response theory (IRT) based methods, the root mean squared difference method (Linn et al., 1981) and the total area method (Rudner, Getson, & Knight, 1980). Specifically, our intent was to compare the items identified by each method and to explain any differences, if possible, in terms of methodological shortcomings or unique features of the methods.

---

[1] This research was supported by a contract from the Air Force Human Resources Laboratory (F33615-84-C-0058). The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Air Force position, policy, or decision, unless so designated by other official documentation.

[2] Laboratory of Psychometric and Evaluative Research Report No. 154. Amherst, MA: University of Massachusetts, School of Education, 1988.

Our interest in the Mantel-Haenszel statistic was easy to justify. Educational Testing Service appears to have considerable interest in this statistic and is presently considering the statistic for wide-scale use in its item analysis and test development work. The statistic can be obtained with relatively simple calculations. A computer program to calculate the statistic would be easy to prepare and inexpensive to use.

The weighted b-value plot method was selected for study because it appeared to provide an answer to the instability problem which limits the usefulness of many of the common IRT and classical item bias statistics (see, for example, Hoover & Kolen, 1984). The weighted b-value plot method incorporates the concept of replication to improve the stability of the item bias detection process. The root mean squared difference and total area methods were chosen for study because they are among the most popular of the IRT-based item bias methods (Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984, 1985). More details on the four methods follow in the next section.

In this paper the terms "differential item difficulty" and "item bias" will be used interchangeably. Although ETS appears to prefer the former term, most researchers using the IRT-based methods studied in the paper seem to prefer the latter. Whichever term is used, the intent is to identify those items in a test where examinees from two groups of interest (e.g. males and females), although similar in ability, differ in their item performance.

Method

Description of the Item Bias Statistics

a. Mantel-Haenszel Statistic

In calculating this item statistic it is necessary first to match the two groups of interest (called "reference" and "focal" in Holland's work, and "majority" and "minority" in item bias research studies) on a relevant criterion to the test under investigation. The most practical criterion is the set of test scores themselves (minus the item under investigation). Modifications such as the test minus items judged on a priori grounds or from preliminary analyses to be potentially biased are also possible. With a K-item test (K + 1 possible scores) and removing the item of interest, it is possible, therefore, to divide the groups into K score groups. In this way, the two groups are matched on a criterion that is related to item performance. Within each test score group (0, 1, ..., K-1) a 2 x 2 contingency table is set up:

		Item j		
		Correct	Incorrect	
Group	Majority (Reference)	A_i	B_i	n_{Ri}
	Minority (Focal)	C_i	D_i	n_{Fi}
	Total	R_i	I_i	n_{+i}

for i = 1,2, ... K = number of matched groups.

A_i, B_i, C_i, and D_i correspond to the numbers of examinees in each of the subgroups. R_i, I_i, n_{Ri}, and n_{Fi} are the marginal totals, and n_{+i} is the combined group total. Considered together, the data available for calculating the Mantel-Haenszel item statistic is 2 x 2 x K.

The Mantel-Haenszel statistic for item j is a weighted sum of the ratio of the odds for answering the item correctly in each score group:

$$MH = \frac{\Sigma W_i \frac{A_i}{B_i} / \frac{C_i}{D_i}}{\Sigma W_i} \qquad [1]$$

$$= \frac{\Sigma W_i A_i D_i / (B_i C_i)}{\Sigma W_i} \qquad [2]$$

where

$$W_i = \frac{B_i C_i}{n_{+i}}. \qquad [3]$$

Substituting Equation [3] into Equation [2] leads to the formula

$$MH = \frac{\Sigma A_i D_i / n_{+i}}{\Sigma B_i C_i / n_{+1}} \quad . \qquad [4]$$

Holland (1985) notes that the MH statistic is "the average factor by which the odds that a reference group member gets item j correct exceeds the corresponding odds for comparable focal group members." The statistic exceeds one when the reference group has the advantage, and is below one when the focal group has the advantage. Associated with the MH statistic is a chi-square test of the hypothesis that all K of the cross products ratios in the 2x2 layers of the 2x2xK table are one. The actual test statistic which is distributed as a chi-square with one degree of freedom is given by Holland (1985).

b. Plot Method

The origin of the plot method for identifying potentially biased items is unknown, though the general approach using non-IRT model parameter estimates was described by Angoff (1982). Hambleton (1982), Hambleton and Murray (1983), and Hambleton, Martois, and Williams (1983) described the plot method in their research papers on goodness-of-fit measures for IRT models. However, it seems likely that Shepard (1981) was the first researcher to describe the general method referred to in this paper as the "plot method."

The advantages of the plot method are that the method (1) provides a basis for comparing item performance in two groups of interest where ability differences are controlled for (item parameter estimates are independent of the groups in which they are obtained), (2) provides a graphical solution for the detection of potentially biased items that is easy for practitioners to understand, (3) recognizes the instability in item bias statistics by focusing only on items which show consistently large differences in item difficulty parameter estimates across a second set of independent samples from the two groups, and (4) provides a sampling distribution of b-value differences under the null hypothesis that there are no differences (this is accomplished by comparing b-value estimates in randomly-equivalent groups). With respect to (4), a cut-off value for interpreting the item bias statistics can be set through the use of the distribution of b-value differences in two randomly-equivalent samples of the same size as the groups to be compared (e.g. Males and Females) (see, for example, Hambleton & Rogers, in press).

There are several additional advantages of the plot method: (1) the problem of sample size is controlled for through the use of baseline plots for interpreting important differences, (2) the baseline plots provide a basis for interpreting the importance of particular independent variables on the invariance property of item difficulty parameter estimates, and (3) the concept of replication replaces the concept of statistical significance testing.

The plot method is not without problems: Considerable computer time is needed to implement the method (a minimum of four LOGIST runs are required), model parameter estimates from the four runs must be equated prior to being compared, and computer plot routines are needed (the alternative is to do the plots by hand, which is tedious). Also, available samples are cut in half which leads to less precisely estimated item parameters. Still, the division of a sample into equal halves allows the researcher to check the replicability of his/her results and generate a sampling distribution of item bias statistics under the correct null hypothesis of no true differences.

The following steps are followed in applying the plot method:

1. Choose the independent variable of interest for the item bias study (e.g., sex, race, geographic region, etc.). Form two groups (e.g., Males and Females) and label them "A" and "B".

2. Count the number of individuals in each group; draw a random sample from the larger group so that both groups (A and B) are of the same size. When the two groups differ in ability, examinees should be sampled from the larger group so that similar ability distributions are obtained. In this way, artifacts in the results due to ability distribution differences can be minimized (Shepard, Camilli, & Williams, 1984).

3. Split both groups randomly to form four equal-sized subgroups (A1, A2, B1, B2).

4. Conduct a three-parameter model analysis on each of the four subgroups (Wood & Lord, 1976) to obtain item and ability parameter estimates.

5. Scale the b-values from each analysis to a mean of zero and a standard deviation of one (or any common mean and standard deviation).

6. Plot the b-values from A1 against A2, and from B1 against B2, to provide baseline information on the amount of scatter to be expected in the parameter estimates due to factors such as sample size and model-data misfit. A1 and A2, and B1 and B2, are randomly equivalent samples.

7. Plot the b-values from A1 and B1, and A2 and B2, to determine if the amount of spread in the plots differs from the baseline plots obtained at step 6. If they do differ, then the independent variable (or a variable confounded with it) is influencing the b-values. A comparison of the A1 and B1, and A2 and B2 plots, permits the researcher to check the replicability of the findings.

8. Plot the differences A1-A2 (the differences in item difficulty estimates in the two samples, A1 and A2) against B1-B2 (the differences in item difficulty estimates in the two samples, B1 and B2) and compare to the plot of A1-B1 against A2-B2. If the plots differ, identify the test items showing consistently large differences in the A and B samples. These items are the ones that may be biased against one of the groups. One useful variation on this step involves plotting the standardized b-value differences. The b value differences are scaled by the standard deviation of the differences to take into account the standard errors of the b values:

a. $$\frac{b(A_1) - b(A_2)}{\sqrt{SE[b(A_1)]^2+SE[b(A_2)]^2}}$$

b. $$\frac{b(B_1) - b(B_2)}{\sqrt{SE[b(B_1)]^2+SE[b(B_2)]^2}}$$

c. $$\frac{b(A_1) - b(B_1)}{\sqrt{SE[b(A_1)]^2+SE[b(B_1)]^2}}$$

d. $$\frac{b(A_2) - b(B_2)}{\sqrt{SE[b(A_2)]^2+SE[b(B_2)]^2}}$$

Then the plot of (a) versus (b) can be compared to the plot of (c) versus (d). Items showing consistently large standardized differences in (c) and (d) are singled out for additional study and identification of possible sources of bias.

There are several variations on the above method. For example, items which on a priori grounds appear to be "biased" can be removed prior to step 4. With ability estimates in hand that are not influenced by potentially biased items, the potentially biased items can be returned to the analysis, and treating the ability estimates as known (fixed), the complete set of item parameter estimates can then be obtained. The variation seems especially useful when the ratio of the number of potentially flawed test items to total test length is high. In this case, the potentially biased test items can "contaminate" the ability estimates and make the overall bias analysis less sensitive. Another variation on the basic method involves estimating ability scores in a combined group analysis and then treating these ability

scores as fixed when calibrating item parameter estimates in the subgroups. The advantage is that simultaneous estimation of abilities and item parameters can be avoided in the smaller samples. This variation was used in the present study.

One shortcoming of the plot-method is that only one item statistic (item difficulty) is used in identifying potentially biased test items. It is possible that the difficulty levels for an item in two groups of interest are equal but the discriminating powers are different. This situation is easily spotted in practice: The item characteristic curves intersect at a point on the ability scale around the item's difficulty level. The plot-method will not identify the item and others like it as potentially biased unless plots are also carried out on the a-value parameter estimates in the two groups of interest. Both the Mantel-Haenszel and the plot-method will have difficulty detecting bias when it results from two ICCs intersecting in the middle range of the ability scores.

c. The Total Area Method

In the "Total Area Method," the area between item characteristic curves for the same item obtained in the two groups of interest over a specified interval on the ability scale is used as an estimate of item bias. The minimum bias (i.e. none) is achieved when the two curves are totally overlapping (i.e. when the items have identical item parameter estimates). Then, the item bias is zero. The more different the two curves over the portion of the ability scale of interest, the larger the area, and the more potentially biased the item is assumed to be. This method of assessing item bias is sensitive to all estimated item

parameters (i.e., b-, a-, and c- parameters). Differences in these parameter estimates in two groups will lead to different curves and thus, to an area of some size between the curves. One important variation on the method is sometimes applied when two ICCs intersect at a point on the ability scale of interest. When they do, the direction of the bias is switched at the point of intersection. In this case, the area representing the bias against each group can be reported in addition to (or instead of) the total area.

d. The Root Mean Squared Difference Method

The Root Mean Squared Difference Method is defined over the same interval on the ability scale as the Total Area Method; however, the square root of the average of the squared differences between the two item characteristic curves at fixed intervals (usually .01) is used as the measure of item bias. Root mean squared difference statistics like the one proposed by Linn et al. (1981) are common in goodness of fit studies. Calculations for the root mean squared difference method and the total area method in this study were carried out on the ability scale between -3 and 3. Prior to computing the item bias statistics, corresponding sets of item statistics were placed on a common scale, using a method described by Linn, et al. (1981).

Description of the Test Data and Examinee Sample

The test data used in the study consisted of the item responses of 937 Cleveland ninth grade students to the first 76 test items (of the 92 items) on the Cleveland Reading Competency Test. The test data were

collected in May of 1985. Item 21 was deleted because of a scoring key problem. There were 451 males and 486 females in the total sample of examinees.

Item parameter and ability estimates obtained from the combined group three-parameter logistic model analysis were treated as "true values." Then, simulated item responses were generated using the three-parameter logistic model to be consistent with the item parameter and ability estimates for the 937 examinees (Hambleton & Rovinelli, 1973; Hambleton & Swaminathan, 1985). The simulated data resembled the original data closely. However, there was no bias in any of the items. These simulated data were used, in part, to provide baseline data for setting cut-off points for interpreting several of the item bias statistics. Supporting evidence for the validity of the simulated data for establishing cut-off points is provided by Hambleton and Rogers (in press) and Rogers and Hambleton (1988).

Procedure

For the purposes of the present study, the differential item performance of males and females on 75 items in the ninth grade Cleveland Reading Competency Test was of central interest. ETS agreed to provide the Mantel-Haenszel chi-square results.

LOGIST '76 (Wood & Lord, 1976) was used in calculating the three sets of item bias statistics based upon item response theory methods.

A modified three-parameter model was used: c-values were set to a value of .25. Fixing the c-values was done because of problems involved in estimating c-parameters (1) with small-sized samples at any time and (2) with the early version of LOGIST we were using (LOGIST '76)

In carrying out the weighted b-value plot method, the total sample (N=937) was divided into female (N=486) and male (N=451) samples, and then the female and male samples were further divided to form two randomly-equivalent female samples (denoted F_1 and F_2) and two randomly-equivalent male samples (denoted M_1 and M_2). Since the male and female ability distributions were nearly equal in size and had similar means and standard deviations (means = .30, .23; standard deviations = .99, 1.15), no examinees were removed from either sample.

Because of the modest sized samples, step four in applying the plot method was revised. First ability estimates were obtained for the total group of examinees using total group item statistics. This step was taken so that improved ability estimates could be obtained due to the increased precision of the item parameter estimates. Next, item parameter estimates were estimated independently in F_1, F_2, M_1, and M_2 (with c = .25) treating the ability estimates obtained in the total group analysis as fixed. The goal was to avoid simultaneous estimation of abililty and item parameters in the small sub-group samples. Finally, the b values in each of the four analyses were scaled

(mean = 0, sd = 1) using common items (items with b values greater than 4.0, were removed from the calculations of means and standard deviations). Standard errors of the b-values were also calculated and placed on the same scale as the rescaled b-values. These calculations and rescalings were carried out using a computer program prepared by the second author. With the b-values on a common scale, the required plots described earlier were obtained.

In addition, LOGIST '76 was run with the combined Female ($F_1 + F_2$) and combined Male ($M_1 + M_2$) samples to obtain two additional sets of item parameter estimates obtained with c = .25, and the same fixed ability estimates described earlier. Again the b-values for each group were scaled to mean=0 and sd=1 using common items (items with b-values greater than 4.0 were removed from the calculations of the means and standard deviations). Then, the a-values were rescaled, and the Total Area and Root Mean Squared Difference statistics were computed. These calculations were carried out on the ability scale over the interval [-3, +3]. A computer program prepared by the second author was again used to obtain the item bias statistics.

One shortcoming of all three IRT-based item bias methods considered in this study was the absence of a criterion for identifying test items showing differences in the groups beyond those that might be expected by chance. The distributions of b-value differences (weighted

and unweighted) for F_1 and F_2 and M_1 and M_2 provide some information about the size of expected random differences. However to standardize the choice of cut-off scores (critical values) across methods for interpreting the item bias statistics, all cut-off scores were set using the simulation results. The bias statistics were calculated from the simulated data and then the distributions of item bias statistics (under the null hypothesis) were calculated. Critical values corresponding to a 1% level of type I errors were set and used to interpret the item bias statistics for the actual test data.

Results

Mantel-Haenszel Method

The Mantel-Haenszel chi-square statistics computed by ETS are reported in the last column of Table 1. Six items appeared to be in need of review because of high Mantel-Haenszel chi-square statistics ($p < .01$). The items were 13, 14, 34, 41, 53, and 60.

Plot Method

Figure 1 shows the plots of weighted b-value differences between the female and male samples in (a) and weighted female-male sample 1 differences and weighted female-male sample 2 differences in (b). The two plots are clearly different. Figure 1a shows essentially no

relationship between the pairs of weighted b-value differences in the Female and the Male samples. A different plot, however, was observed in Figure 1b. If sex, or a variable confounded with sex was not a factor in item performance, Figure 1b would show a similar pattern to Figure 1a. The differences between the first and second female and male samples showed some consistency. Items that showed consistent differences were identified as potentially biased.

The simulation results reported in Figure 2 showed that, for weighted b-value differences, a cut-off score of 1.40 would be appropriate for holding type I errors to a level of 1%. Figures 2a, 2b, and 1a were similar and highlighted (1) the proper shapes of the plots when bias is not operating and (2) the applicability of computer simulation techniques for obtaining cut-off scores. With a criterion of $\pm$ 1.40, eight test items (see Table 1) showed a consistent difference that large in the female and male samples: 2, 10, 13, 34, 46, 53, 60, and 75.

Total Area Statistic

The IRT Total Area statistics for the 75 test items with the Female and Male groups are reported in Table 1. With a cut-off score of .50 applied to the Female and Male sample Total Area Statistics, seven test items were identified: 12, 13, 25, 34, 46, 68, and 73.

Root Mean Squared Difference Statistic

The Root Mean Squared Difference statistics for the Combined Female and Combined Male groups are also reported in Table 1. The critical value for interpreting these statistics is .11 (obtained from

the simulation results). Five test items exceeded the critical value: Items 13, 25, 34, 46, and 73.

Comparison of Methods

Table 2 provides a complete list of the potentially biased test items identified in the item bias analyses. The number of items identified varied from 5 to 8 across the four methods. Several measures of agreement were available. Rank order correlations were not used because the plot method does not lead to a simple ranking of items like the other three methods. In addition, of central interest was the level of agreement among the methods in identifying the worst or most (potentially) biased test items. Table 3 provides measures of agreement among the four methods. In the upper portion of the matrix, the method with the lowest number of potentially biased items served as the denominator; in the lower portion of the matrix, the method with the largest number of potentially biased items served as the denominator.

The statistics in Table 3 suggested that there was moderate agreement about the items identified as potentially biased by the four methods. The single exception was the Total Area and RMSD methods, which led to very similar results.

The appropriate question at this point seemed to concern the reasons for the different results: Were the differences due to methodological problems associated with the methods or to unique features of the methods? The methodological problems included:

1. Imprecise placement of the cut-off score. This problem could lead to an over- or underdetermination of potentially biased test items.

2. Type I errors. This problem could lead to items being labelled as "potentially biased" due solely to random errors.

3. Poorly estimated IRT item parameters. This problem would lead to unstable item bias statistics.

The first problem seemed to apply to the Weighted b-Value Plot Method. The problem was most likely to affect this method because a cut-off score must be set to identify the 1% significance level for a bivariate distribution of variables. The placement of a cut-off point in a region of very limited data in a bivariate distribution is difficult and apt to be quite imprecise. Also, the simulated distributions (see Figure 2a and 2b) which were used in setting the cut-off point, were somewhat more homogeneous than the plot obtained with real data from randomly equivalent groups (see Figure 1a). The result was that the cut-off point may have been underestimated. Items 2, 10, and 75 were borderline and were identified with only the Weighted b-Value Plot Method. If a more precise placement of the cut-off score had been possible, these items would not likely have been identified.

Three other items, 12, 14, and 41, were identified by only one of the methods. Also, they were close to the cut-off point. Since, on the average, one item per method would be expected to be identified as potentially biased because of type I errors, it is probable that those three items were on the list in Table 2 because of type I errors. Therefore, it would appear that six of the fourteen items identified as

potentially biased were probably misclassified because of problems associated with placing the cut-off scores or because of type I errors.

The third problem almost certainly applied to items 25, 68, and 73. Table 3 provides the item parameter estimates for these items and the other items identified as potentially biased. The b-value estimates for items 25, 68, and 73 were large and unstable. Unstable b-values for items with poor discriminating power are a well-known occurrence when the 1976 version of LOGIST is used in parameter estimation. The point biserials for these three items were quite low (below .15 for all three items). The estimation of item parameters is affected by estimation procedures, sample size, and item quality. A review of the statistics and content for these three items suggested strongly that the estimation problems were due to error-filled data resulting from poorly prepared distractors and/or multiple correct answers.

When items 2, 10, 12, 14, 25, 41, 68, 73, and 75 were removed from Table 2 because of (1) the imprecision in the cut-off score placement, (2) type I errors, and/or (3) parameter estimation errors, high levels of agreement were obtained for items identified as potentially biased across methods. Of the five remaining items, two (13 and 34) were identified by all four methods. A third item was identified by the three IRT methods (item 46) but <u>not</u> by the Mantel-Haenszel method. The two remaining items (item 53 and 60) were only identified by the Weighted b-Value Plots Method and the Mantel-Haenszel Method. Female and male ICCs for items 46, 53, and 60 are presented in Figures 3, 4, and 5, respectively.

There are two unique features associated with the methods that can result in different items being identified:

1. The Mantel-Haenszel statistic in Equation [4] and the Weighted b-Value Plot Method cannot identify bias when the bias results from intersecting item characteristic curves.

2. The choice of interval over which bias is determined affects the Total Area and RMSQ item bias statistics.

Item 46 was identified by the three IRT methods, but not the Mantel-Haenszel statistic. A plot of the female and male ICCs (see Figure 3) showed that differential performance between the two groups began to occur only at ability levels above 1.0. Since the mean ability level of the males was .30 and the females was .23, probably the MH statistic did not identify the bias because the differences were obtained in a region on the ability scale beyond which most examinees scored. The IRT-based methods, by examining the whole ability range (or at least -3.0 to +3.0), identified item 46 as biased.

Items 53 and 60, on the other hand, were identified as biased by the MH statistic and the Weighted b-Value Plot Method, but not by the other two IRT-based methods. The ICC plots which are very similar are shown in Figures 4 and 5, respectively. One possible explanation is that item performance differences in the males and females were substantial over the region of the test score scale where most examinees scored. Therefore, the MH statistic would be very sensitive to the performance differences. Also, the b-value differences in the male and females were substantial and consistent over the two samples. On the other hand, the IRT Total Area and RMSQ methods considered the differences across a wider range of ability scores, and over the full

range of scores considered, the differences were not large enough to identify the two items as potentially biased differences. A different result would likely have been obtained if the interval over which bias was defined was revised from [-3 to 3] to [-2 to 2].

Conclusions

The evidence seems clear that the four methods lead, methodological problems aside, to nearly the same set of potentially biased test items. The real differences that were observed among the methods due to unique features of the methods produced only a small number of disagreements. The four methods did not differ in their detection of potentially biased items by more than an item or two after methodological shortcomings were taken into account. The choice of interval over which bias was defined appeared to be the cause of the three differences that were found.

While the results from a single study do not warrant strong conclusions, further research on the Mantel-Haenszel statistic seems highly worthwhile. As predicted by Holland (1985), the Mantel-Haenszel statistic provided quick cheap empirical estimates of differential performance that, at least for moderate sample sizes, were as good or better than several of the popular IRT-based estimates. The IRT-based methods produced similar results to the MH statistic, but they were time-consuming to carry out and the costs of the analyses, especially for the Weighted b-Value Plot Method, were considerably higher.

One caution we note about the Mantel-Haenszel statistic is that it is "sample dependent," since the odds ratio for each score level used

in the calculations of the Mantel-Haenszel statistic is weighted by a factor reflecting the number of examinees at each score level. In different samples, where the distributions of abilities are different, the statistic might not produce the same result. The advantages and disadvantages of "sample dependent" and "sample independent" item bias statistics will need to be addressed when selecting an item bias statistic for each application. The same information must be kept in mind when interpreting item bias statistics. The "unweighted" IRT-based methods which were used in this study, on the other hand, compare ICCs over a full range of ability without regard to the number of examinees at each score level.

References

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

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Hambleton, R.K. (1982). Applications of item response models to NAEP mathematics exercise results. Final Report -- ECS Contract No. 02-81-20319. Denver, CO: Educational Commission of the States.

Hambleton, R.K., Martois, J.S., & Williams, C. (1983, April). Detection of biased test items with item response models. Paper presented at the annual meeting of AERA, Montreal.

Hambleton, R.K., & Murray, L.N. (1983). Some goodness of fit investigations for item response models. In R.K. Hambleton (Ed.) Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Hambleton, R.K., & Rogers, H.J. (1986). Evaluation of the plot method for identifying potentially biased test items. In S.H. Irvine, S. Newstead, & P. Dann (Eds.), Computer-based human assessment. Hingham, MA: Kluwer-Nijhoff.

Hambleton, R.K., & Rogers, H.J. (in press). Promising directions for assessing item response model fit to test data. Applied Psychological Measurement.

Hambleton, R.K., & Rovinelli, R.J. (1973). A Fortran IV program for generating examinee response data for logistic test models. Behavioral Science, 18, 74.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Hingham, MA: Kluwer-Nijhoff.

Holland, P.W. (1985). On the study of differential item difficulty. Princeton, NJ: Educational Testing Service.

Holland, P.W., & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Technical Report No. 86-31. Princeton, NJ: Educational Testing Service.

Hoover, H.D., & Kolen, M.J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

AERA in 86, Bias.1

Tronson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias in a test of comprehension. Applied Psychological Measurement, 5, 159-173.

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

Rogers, H.J., & Hambleton, R.K. (1988). Evaluating computer simulated baseline statistics for interpreting item bias statistics. Laboratory of Psychometric and Evaluative Research Report No. 162. Amherst, MA: School of Education, University of Massachusetts.

Rudner, L.M., Getson, P.P., & Knight, D.L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Shepard, L.A. (1981). Identifying bias in test items. In B.F. Green (Ed.), Issues in testing: coaching, disclosure and ethnic bias. San Francisco: Jossey-Bass.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-138.

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

Wood, R.L., & Lord, F.M. (1976). A user's guide to LOGIST. Research Memorandum. Princeton, NJ: Educational Testing Service.

AERA in 86, Bias.1

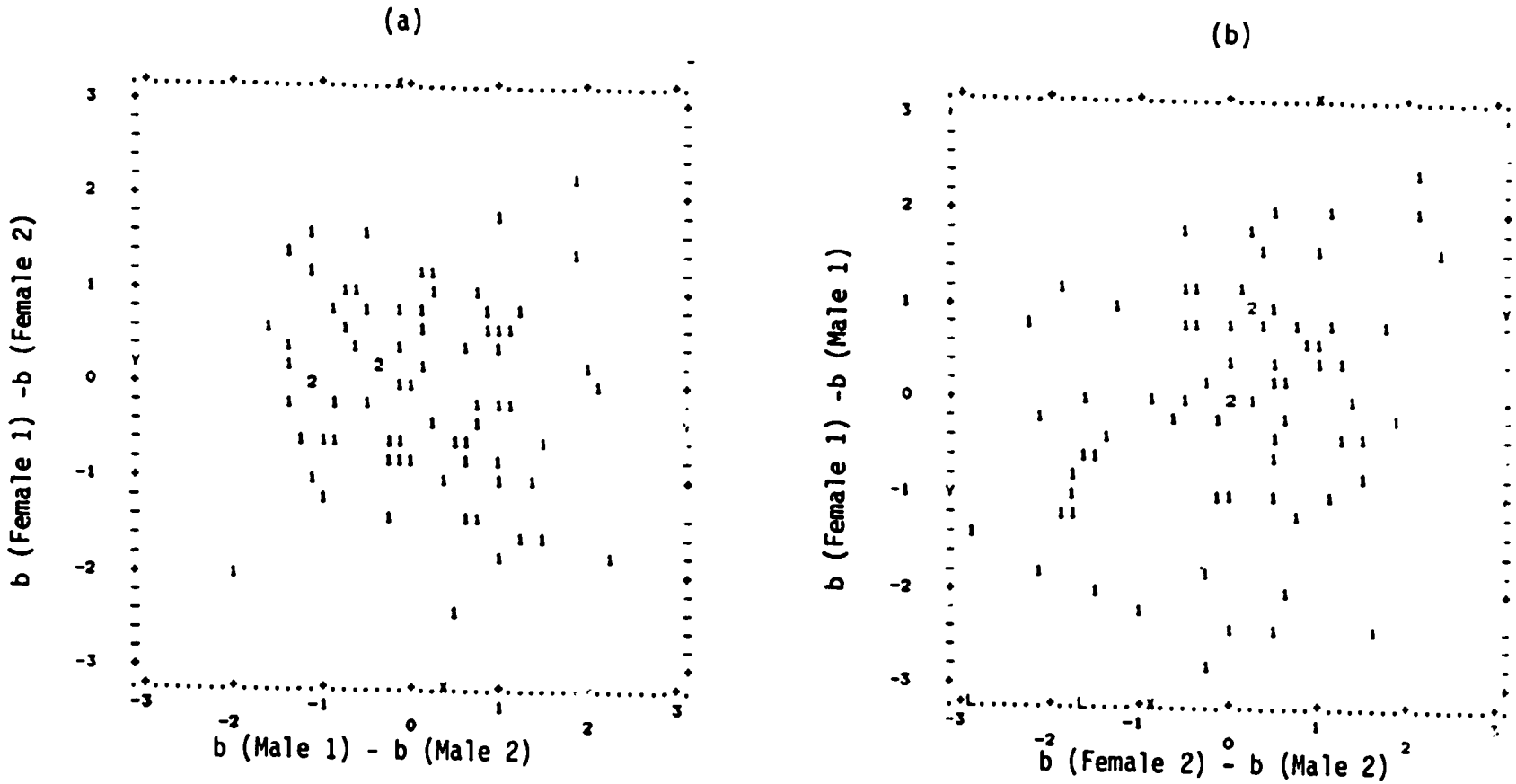

Figure 1. Plots of weighted b value differences, (Female Sample 1 - Female Sample 2) versus (Male Sample 1 - Male Sample 2) in (a) and (Female Sample 1 - Male Sample 1) versus (Female Sample 2 - Male Sample 2) in (b).

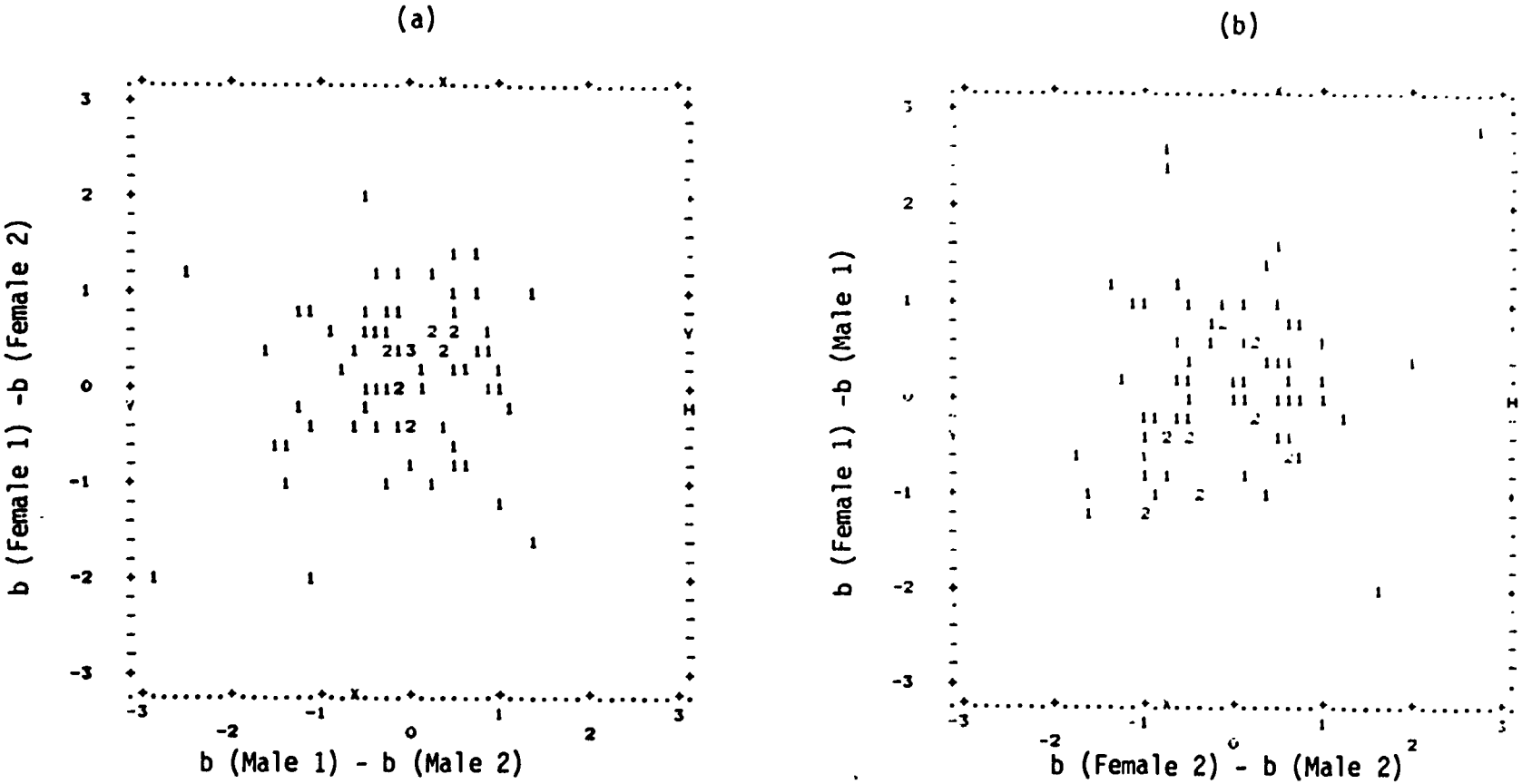

Figure 2. Plots of weighted b value differences, simulated data, (Female Sample 1 - Female Sample 2) versus (Male Sample 1 - Male Sample 2) in (a) and (Female Sample 1 - Male Sample 1) versus (Female Sample 2 - Male Sample 2) in (b).

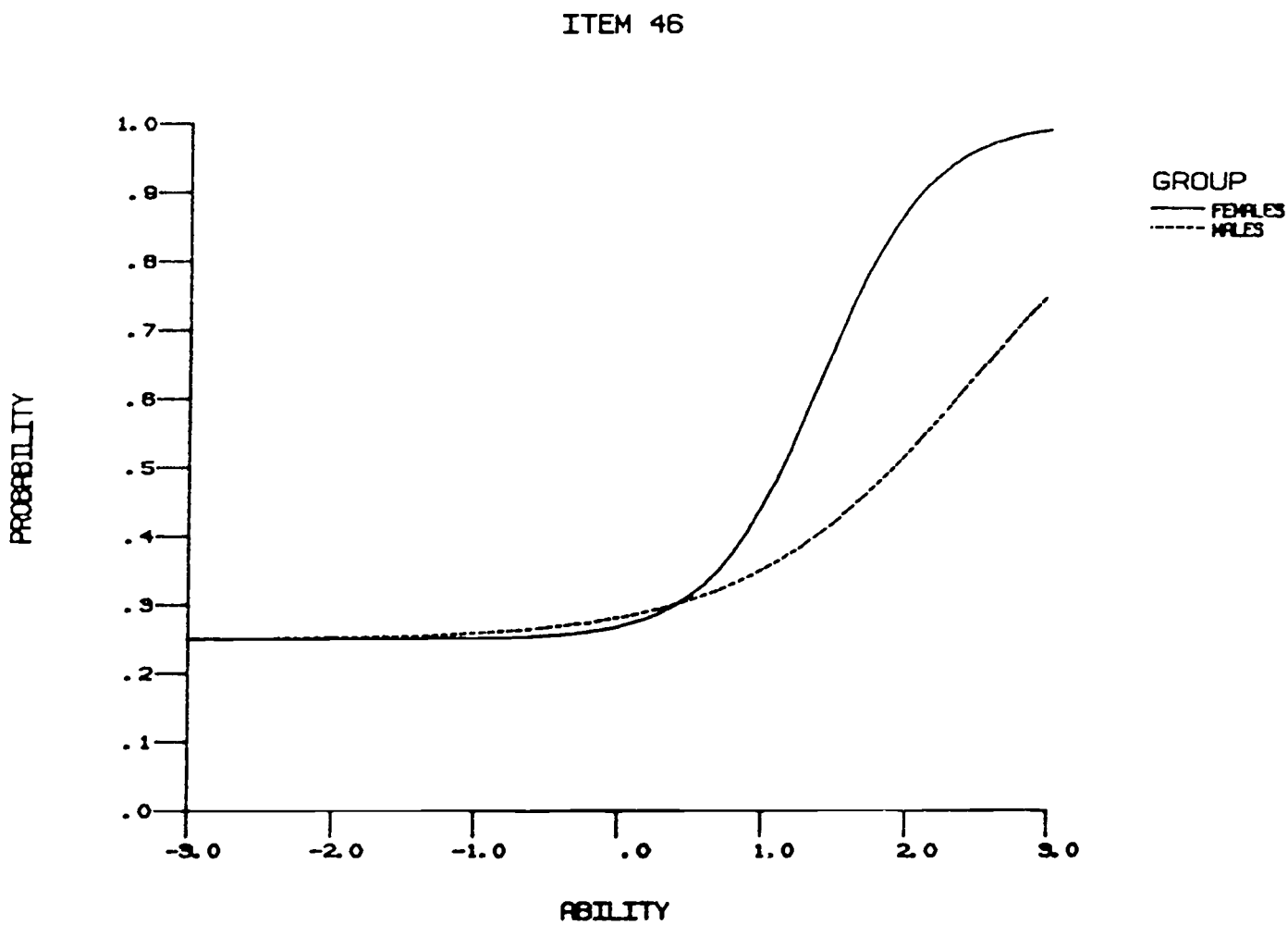

Figure 3. Item 46 ICCs for females and males.

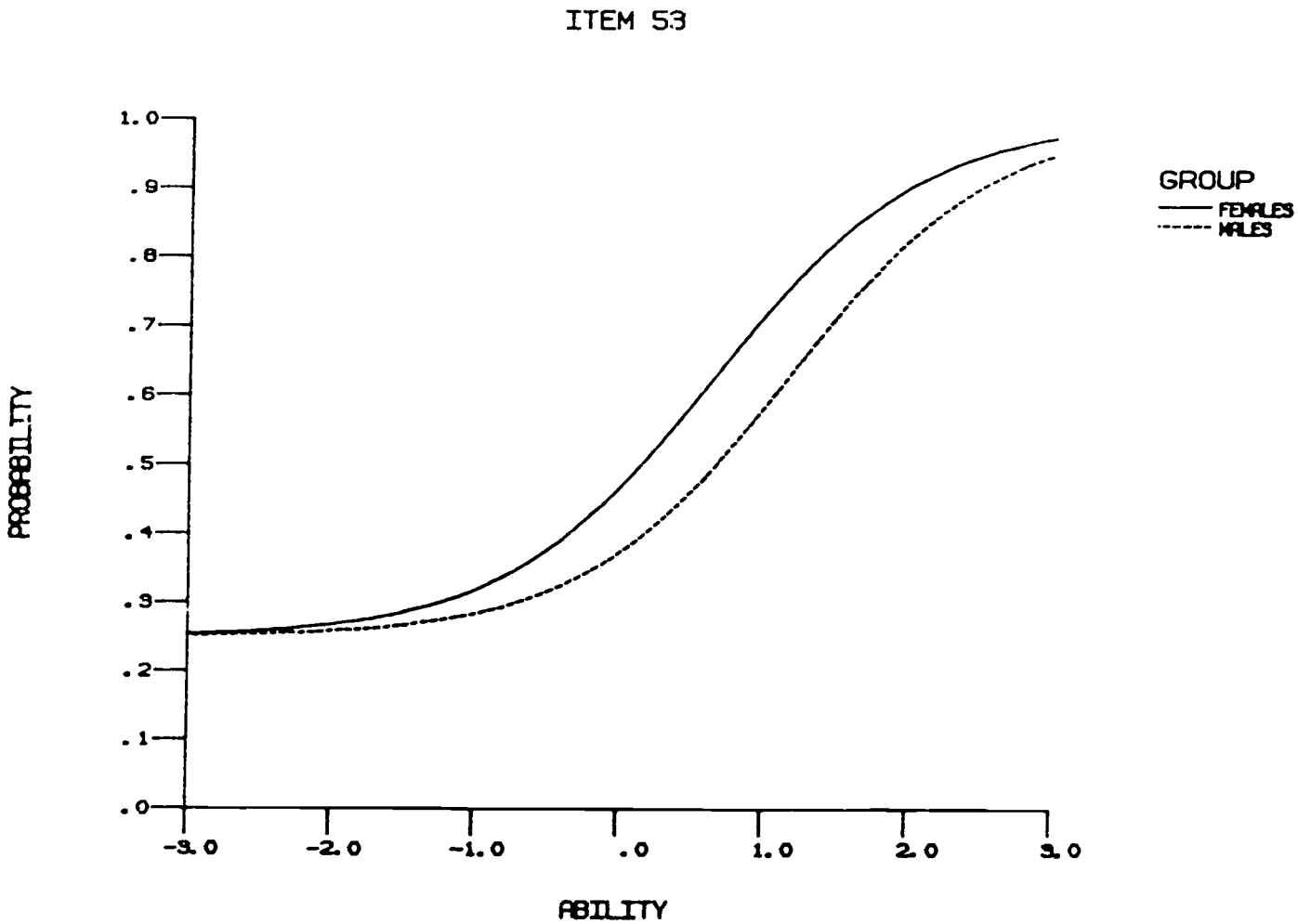

Figure 4. Item 53 ICCs for females and males.

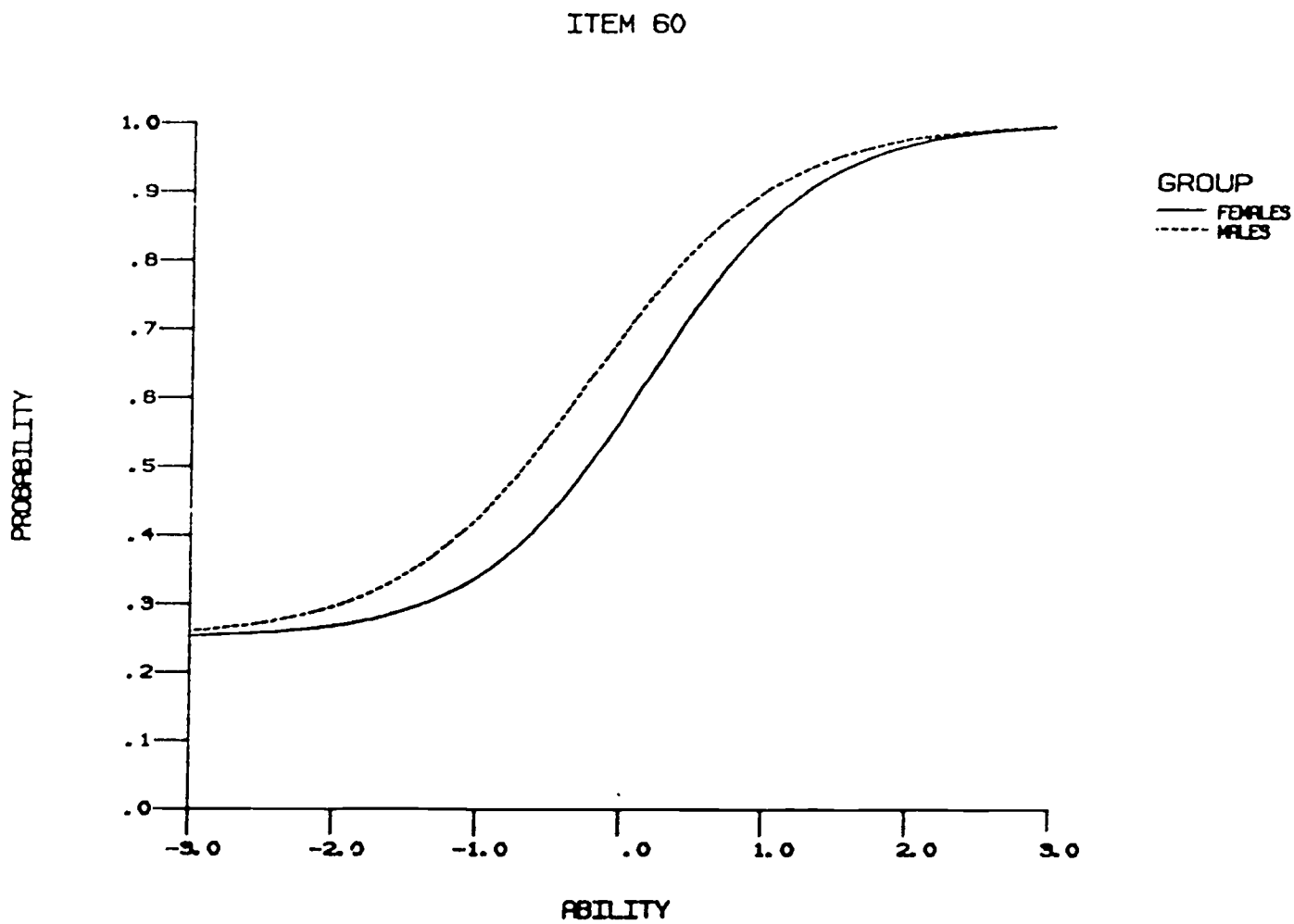

Figure 5. Item 60 ICCs for females and males.

Table 1 (continued)

Test Item	Weighted b-Value Difference Sample 1 (F1, M1)	Weighted b-Value Difference Sample 2 (F2, M2)	Total Area (F, M)	Root Mean Squared Difference (F, M)	Mantel-Haenszel X^2 Statistic (F, M)
31	1.76	1.11	.26	.053	1.90
32	.06	2.38	.23	.044	2.73
33	- .45	- .39	.09	.019	.02
34	1.67	3.89	.52	.129	15.51*
35	1.66	.03	.37	.080	.73
36	.47	- .70	.17	.037	.01
37	- .41	- .84	.15	.038	.35
38	2.27	- .77	.17	.039	.12
39	.48	-1.73	.13	.032	.03
40	-1.32	.09	.11	.024	1.49
41	.24	1.77	.33	.067	6.67*
42	- .76	- .75	.29	.056	1.91
43	- .62	- .25	.25	.048	5.11
44	-1.15	-2.04	.27	.053	2.68
45	1.63	.51	.22	.044	2.20
46	-2.17	-2.43	.60	.168	2.87
47	-1.24	- .32	.19	.038	1.15
48	-1.00	- .60	.21	.041	1.06
49	-1.58	2.50	.09	.017	1.01
50	.69	.12	.18	.033	.08
51	- .18	-1.13	.23	.044	.63
52	-1.91	.18	.16	.039	.57
53	-2.17	-1.91	.37	.076	9.49*
54	-1.49	.86	.06	.015	2.08
55	- .67	.28	.26	.047	1.63
56	1.89	-1.21	.33	.060	.87
57	- .44	- .20	.09	.019	.35
58	.06	- .49	.16	.034	.05
59	.42	-1.12	.09	.017	.44
60	2.17	1.76	.30	.065	7.31*
61	- .42	-1.67	.19	.040	2.37
62	1.71	.94	.30	.066	.32
63	- .24	-1.78	.22	.054	2.85
64	-1.50	.50	.18	.037	2.03
65	-1.05	-1.55	.31	.054	2.67

- Continued -

Table 1 (continued)

Test Item	Weighted b-Value Difference Sample 1 (F1, M1)	Weighted b-Value Difference Sample 2 (F2, M2)	Total Area (F, M)	Root Mean Squared Difference (F, M)	Mantel-Haenszel X^2 Statistic (F, M)
66	-1.29	.49	.09	.018	.06
67	- .22	-1.04	.18	.043	.89
68	- .23	.02	.55	.091	.01
69	-1.07	.90	.30	.058	.91
70	.85	- .04	.24	.043	.66
71	1.34	.37	.11	.025	2.86
72	.37	-.71	.04	.008	.05
73	.08	.94	.69	.126	5.54
74	- .56	2.35	.17	.036	1.46
75	-2.34	-1.59	.42	.082	6.46

*p < .01 (Critical value = 6.64)

Table 2

Potentially Biased Test Items
Identified With Each Method (Labelled "X")

Test Item	Weighted b-value Differences ±(1.40)	Total Area F vs. M (.50)	RMSQ F vs. M (.11)	Mantel-Haenszel X^2 Statistic
2	X			
10	X			
12		X		
13	X	X	X	X
14				X
25		X	X	
34	X	X	X	X
41				X
46	X	X	X	
53	X			X
60	X			X
68		X		
73		X	X	
75	X			
Total	8	7	5	6

*Cut-off scores are in parentheses.

Table 3

Agreement Between Pairs of Item Bias Statistics[1]

Method	Method 1 (8)	2 (7)	3 (5)	4 (6)
1. Weighted b-value Differences	-	42%	60%	67%
2. Total Area	37.5%	-	100%	33%
3. RMSD	37.5%	70%	-	40%
4. Mantel-Haenszel X^2 Statistic	50%	29%	33%	-

[1]The number of potentially biased test items for each method is in parentheses.

Table 4

Summary of Item Statistics
for Potentially Biased Items*

Item	Female Samples						Male Samples					
	1		2		Total		1		2		Total	
	b	a	b	a	b	a	b	a	b	a	b	a
2	1.14	.80	1.14	.84	1.18	.79	.52	1.10	.78	.82	.66	.87
10	.13	1.49	- .10	.94	.02	1.04	- .17	.76	- .81	.45	- .42	.59
12	.36	.67	.58	1.52	.49	.94	- .13	.61	.28	.50	.04	.50
13	.30	1.25	.58	1.41	.44	1.22	- .13	1.79	- .46	.47	- .30	.74
14	1.40	1.97	1.55	1.42	1.54	1.57	1.64	.76	1.18	1.61	1.41	1.04
25	19.05	.05	8.89	.14	11.34	.10	6.20	.22	11.99	.12	8.08	.17
34	- .57	1.21	- .33	1.71	- .46	1.42	- .99	1.14	-1.28	1.03	-1.17	1.04
41	1.92	1.04	2.41	.89	2.22	.98	1.87	.68	1.69	1.18	1.86	.79
46	1.29	2.13	1.38	1.14	1.40	1.55	2.67	.60	2.08	1.20	2.46	.75
53	.57	.65	.70	1.23	.65	.82	1.16	.80	1.08	.97	1.16	.83
60	.21	1.29	.14	.85	.18	1.00	- .13	1.35	- .31	.70	- .22	.90
68	2.62	.09	59.65	.01	32.03	.01	12.18	.02	47.91	.01	39.38	.01
73	2.89	.32	3.94	.33	3.58	.30	2.88	1.58	2.47	1.61	2.76	1.73
75	.44	.57	.65	.70	.55	.56	1.16	.70	1.26	.40	1.28	.49

* The c-parameter in the three-parameter logistic model was set to a value of .25 for all analyses.