DOCUMENT RESUME

ED 378 197 TM 022 527

AUTHOR Tang, Huixing
TITLE A New IRT-Based Small Sample DIF Method.
PUB DATE Jan 94
NOTE 19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 27-29, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Grouping; *Analysis of Variance; Goodness of Fit; Identification; *Item Bias; *Item Response Theory; Reference Groups; Robustness (Statistics); *Sample Size; Sampling; Simulation; Test Construction
IDENTIFIERS Calibration; Mantel Haenszel Procedure; Polytomous Scoring; Power (Statistics); *Residual Scores

ABSTRACT

This paper describes an item response theory (IRT) based method of differential item functioning (DIF) detection that involves neither separate calibration nor ability grouping. IRT is used to generate residual scores, scores free of the effects of person or group ability and item difficulty. Analysis of variance is then used to test the group differences in residual scores. Simulation studies were conducted to examine the power and error rate of the method in terms of varying sample sizes, difference reference or focal group ratios, and varying degrees of group ability difference. Also investigated was the robustness of the procedure to moderate model-data misfit. Mantel-Haenszel statistics were also computed for comparison purposes. Desirable features of the new (IRT-ANOVA) method include: (1) concurrent calibration; (2) no ability grouping; (3) capability to examine interaction effects; (4) applicability to both dichotomous and polytomous items; and (5) applicability to relatively small sample sizes. Limitations and advantages of the method are discussed. Six figures illustrate the analyses. (Contains 2 references.) (SLD)

A New IRT-Based Small Sample DIF Method

Huixing Tang

The Psychological (poration

Paper Presented at the Annual Meeting of the
Southwest Educational Research Association
January 27-29, 1994, San Antonio, Texas

I. OBJECTIVE

IRT-based DIF methods typically involve ability grouping and/or a separate calibration for each of the groups being compared. One limitation of these methods is that they are often inapplicable in situations where the sample size, particularly that of the focal or minority group, is small. This is because both separate calibration and ability grouping require relatively large sample sizes. A frequent alternative in small sample situations is to use the Mantel-Haenszel(MH) procedure generally considered 'particularly useful when groups are small', and a 'uniformly most powerful and unbiased test...'(Hills, 1989, p. 9).

This paper describes an IRT-based DIF method which involves neither separate calibration nor ability grouping. IRT is used to generate residual scores: scores which are free of the effects of person (or group) ability and item difficulty. Analysis of variance is then used to test the group differences in residual scores. Simulation studies were conducted to examine the power and error rate of the method in terms of varying sample sizes, different reference/focal group ratios, and varying degrees of group ability difference. Also investigated was the robustness of the procedure to moderate model-data misfit. MH statistics were also computed for comparison purposes. Since the new method employs both IRT and ANOVA, it will be henceforth referred to as the IRT-ANOVA method.

II. THE PROCEDURE

The IRT-ANOVA method consists of the following steps:

1. Calibrate the data with an appropriate IRT model to obtain estimates of person ability and item difficulty, or step difficulty for polychotomously scored items.

2. Using the estimates obtained in step 1, compute P_{ij}, the probability of person i responding correctly to item j, or P_{ijk}, the probability of person i scoring in the kth category of item j.

3. Using the probabilities obtained in step 2, compute E_{ij}, the expected score for person i on item j. For dichotomous items, P_{ij} is equivalent to the expected score. For polychotomous items with $m+1$ categories, E_{ij} is computed by

$$E_{ij}=\sum_{k=0}^{m_i} kP_{ijk}$$

4. For each person, the residual item score, R_{ij}, is computed by subtracting E_{ij} from the observed item response X_{ij}.

$$R_{ij}=X_{ij}-E_{ij}$$

5. Perform analysis of variance or regression analysis on the data with R_{ij}'s used as values of the dependent variable and DIF factor(s) (e.g. gender, race) under investigation as the independent variable(s). The resulting F ratio is used as the test statistic for DIF.

The most salient feature of the method is that it uses IRT to control for group ability difference and employs familiar inferential procedures to test the DIF effect. Residuals can be construed as item scores corrected for person ability and item difficulty. They are expected to be random (i.e., error) with a mean of 0. A positive residual may imply that the person is scoring higher than expected based on overall test performance. A negative residual may imply that the person is scoring lower than expected. Consistently high (or low) residual values for a particular subgroup may imply that the item favors (or disfavors) this subgroup.

The method has several desirable features. First, it is capable of processing two or more DIF factors simultaneously. This makes it possible to examine interaction effects between DIF factors or to examine main effect DIF by controlling for confounding variable(s). These are areas yet to be explored in DIF analysis in general. Second, with concurrent calibration, the procedure bypasses the problem of scale shift due to separate calibrations. Third, since it does not involve ability grouping, there is no loss of within-ability-interval information or arbitariness in deciding on the number of intervals to be used. Fourth, it is applicable to both dichotomously and polychotomously scored items. And lastly, because the method does not involve a separate calibration, it is applicable to relatively small focal group sample sizes.

III. THE DESIGN OF SIMULATION STUDIES

Simulation studies were conducted to examine the power and error rate of the method in comparison with the MH method. The DIF factor simulated was gender, with males being the reference group and females the focal group. Four conditions were simulated: the reference and the focal groups are (1) equal in sample size and ability, (2) unequal in sample size (67% males and 33% females) but equal in ability, (3) equal in sample size but unequal in ability (the mean logit is .6 higher for males than for females), and (4) unequal in both sample size and ability. The examinee sample is normally distributed and ranges from 200 to 1200 in increments

of 200. Each condition is replicated 300 times for each sample size level.

Item characteristics for 30 items were generated using the unit normal distribution. They were randomly generated for each replication so that the effect of item difficulty on DIF was controlled. Four DIF items were introduced. Two of them favored the reference group, and the other two favored the focal group, both by .6 in logit difficulty. Also, two of the DIF items and one non-DIF item were simulated as moderately misfitting items (with mean square infit values around 1.10). Their inclusion allows for the examination of the effect of lack of unidimensionality on DIF detection using the current method. These items were generated using a secondary ability distribution with a correlation of .5 with the primary ability distribution that generated the rest of the items. It should be noted that, due to imperfect correlation between the two distributions and hence the effect of regression toward the mean, the secondary distribution invariably favors the low ability group and disfavors the high ability group under conditions 3 and 4 mentioned above. This is analogous to a situation in which a second dimension favors the low ability group on a subset of items. Figure 1 presents a graphic display of the combinations of items, sample sizes, and the conditions.

In this study, the Rasch model was used for both data simulation and calibration, though there is nothing in the design of the procedure that prevents the use of other IRT models. Only dichotomous items and one DIF factor are used, which makes it possible to compare with the results from the MH procedure. A computer program was written by the author which processes data simulation, Rasch calibration using the unconditional maximum likelihood method, and computation of ANOVA and other relevant statistics.

VI. FINDINGS AND DISCUSSIONS

The power of the IRT-ANOVA method when data fit the model

Power is defined as the percent of correct rejections or rejection rate for DIF items. The significance level is set at .05. Figure 2 displays the rejection rate for the two non-misfitting DIF items under each condition. It shows that the rejection rate under conditions 1 and 3 (equal group size), is generally higher than under conditions 2 and 4 (unequal group size), respectively. The differences, ranging from 0 to 7 percent, tend to decrease as the sample size increases.

The effect of group size differences on power implies that a larger sample size is needed when the group sizes are unequal than when they are equal. In the two

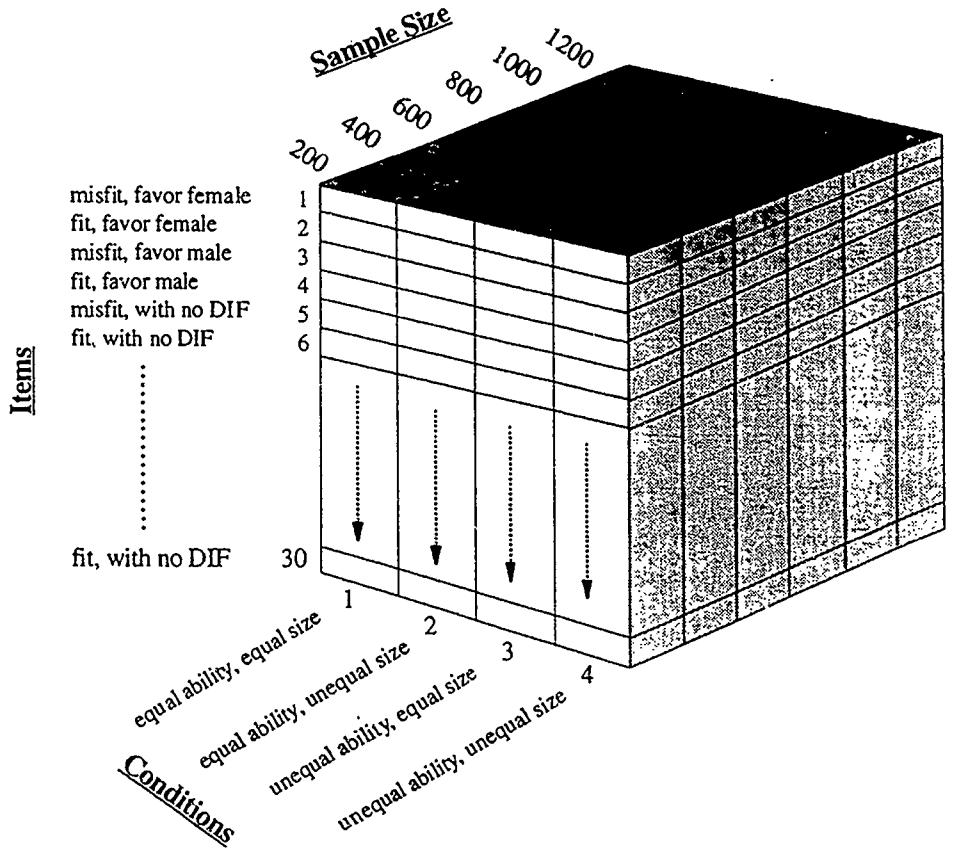


Figure 1. Design of the Simulation Studies

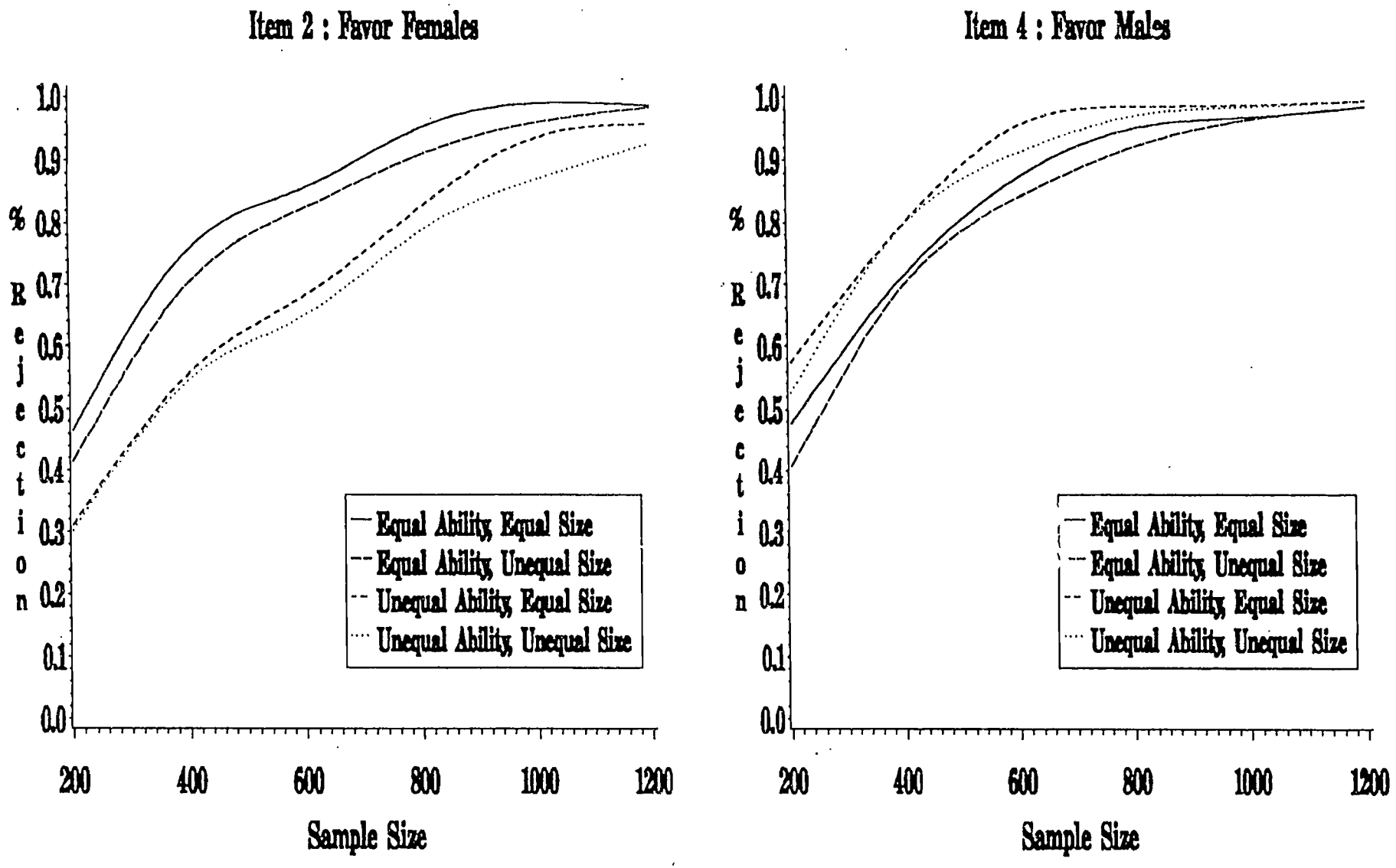


Figure 2. Percent of rejection of DIF Items under four conditions when data fit the model

group case, it can be shown mathematically that, given a fixed total sample size and equal variances for the two groups, the power of the test decreases as the group size difference increases (Cohen, 1969).

The effect of group ability difference on power exhibits a more complex relationship. For Item 2, which favors females (the low ability group under conditions 3 and 4), the rejection rate is lower when there is an ability group difference than when there is no difference. For Item 4, which favors males (the high ability group), the rejection rate is higher when there is an ability group difference. In both situations, the differences between conditions 1 and 2 on the one hand and conditions 3 and 4 on the other decrease as the sample size increases.

The cause(s) of the interaction between group ability difference and the direction of DIF (whether it favors the low or high ability group) are not clear. The reduced power for item 2 could be due to the underadjustment of group ability difference. The advantage of the low ability group on this item may be adversely affected by the underadjustment. The increased power of item 4 could also be caused by underadjustment of group ability difference. But in this case, the advantage of the high ability group on the item is compounded by the underadjusted difference.

The effect of group ability difference implies that DIF magnitude should be interpreted in light of the magnitude of the group ability difference and the direction of DIF. DIF favoring a high ability group is more easily identified than DIF favoring the low ability group using the current method.

The power of the IRT-ANOVA method when data do not fit the model

Figure 3 displays the rejection rate of the two moderately misfitting items. For item 1, which favors the low ability group, the rejection rate is higher under conditions 3 and 4 than under conditions 1 and 2. For item 3, which favors the high ability group, the rejection rate is lower under conditions 3 and 4 than under conditions 1 and 2. Recall that the responses to the misfitting items are simulated using a secondary ability distribution that favors the low ability group. Consequently, the rejection rate increases when the item favors the low ability group and decreases when the item favors the high ability group.

The interaction between model-data misfit and the direction of DIF suggests that DIF statistics ought to be interpreted in light of item fit statistics. With the type of misfit investigated in this study, DIF favoring the low ability group is easier to identify than DIF favoring the high ability group. A different relationship may

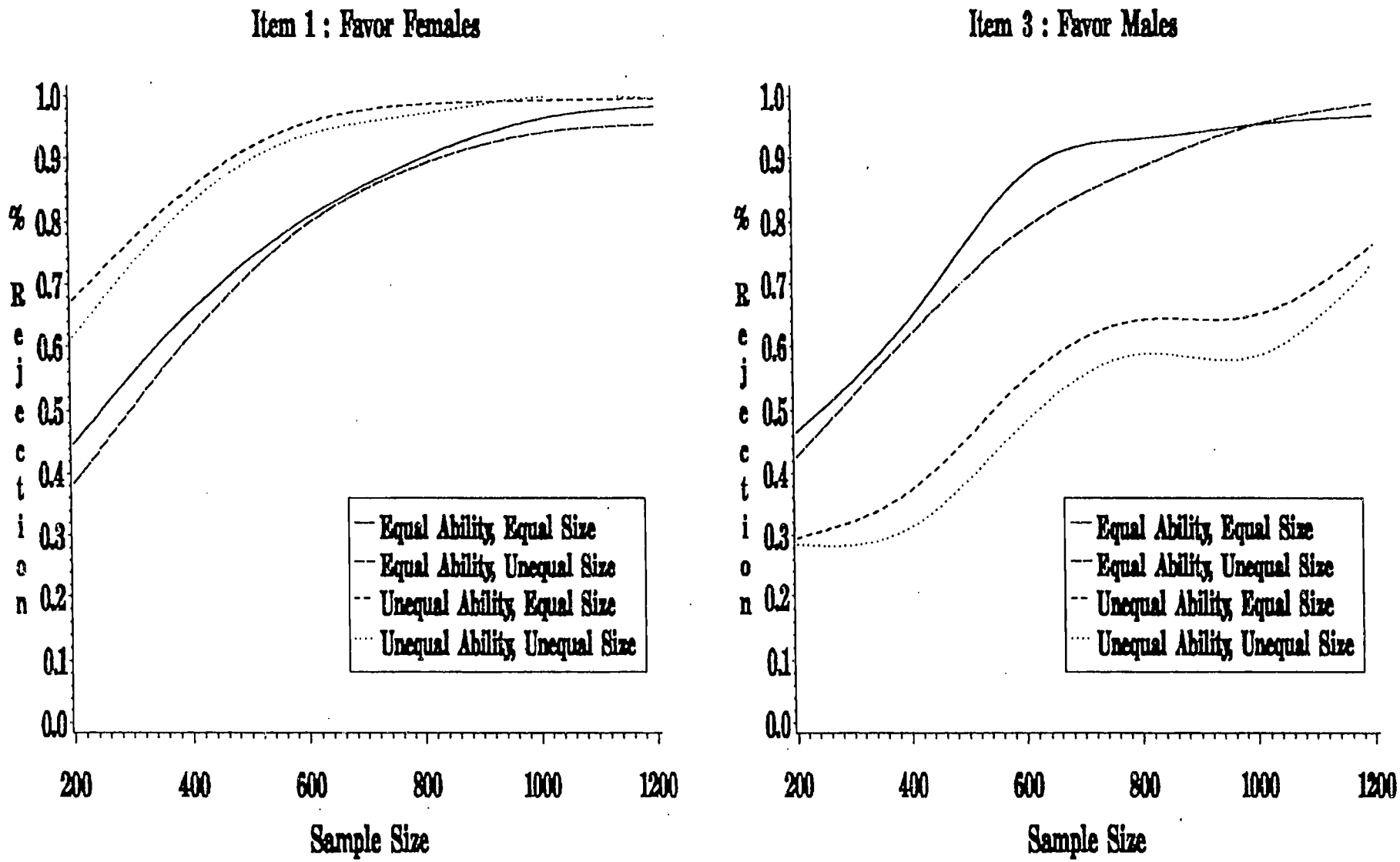


Figure 3. Percent of rejection of DIF items under four conditions when data do not fit the model

occur with a different type of misfit. If, for example, the second dimension favors the high ability group, then DIF favoring the high ability group will be easier to detect than DIF favoring the low ability group.

Figure 4 displays the rejection rate of the four DIF items under conditions 1 and 2. It shows that, when there is no group ability difference, the rejection rate for the non-misfitting items is generally higher than that for the misfitting items. But the difference is small for the most part. The relatively small effect could be due to the fact that the impact of the second dimension is evenly distributed between the two groups whey they do not differ in ability.

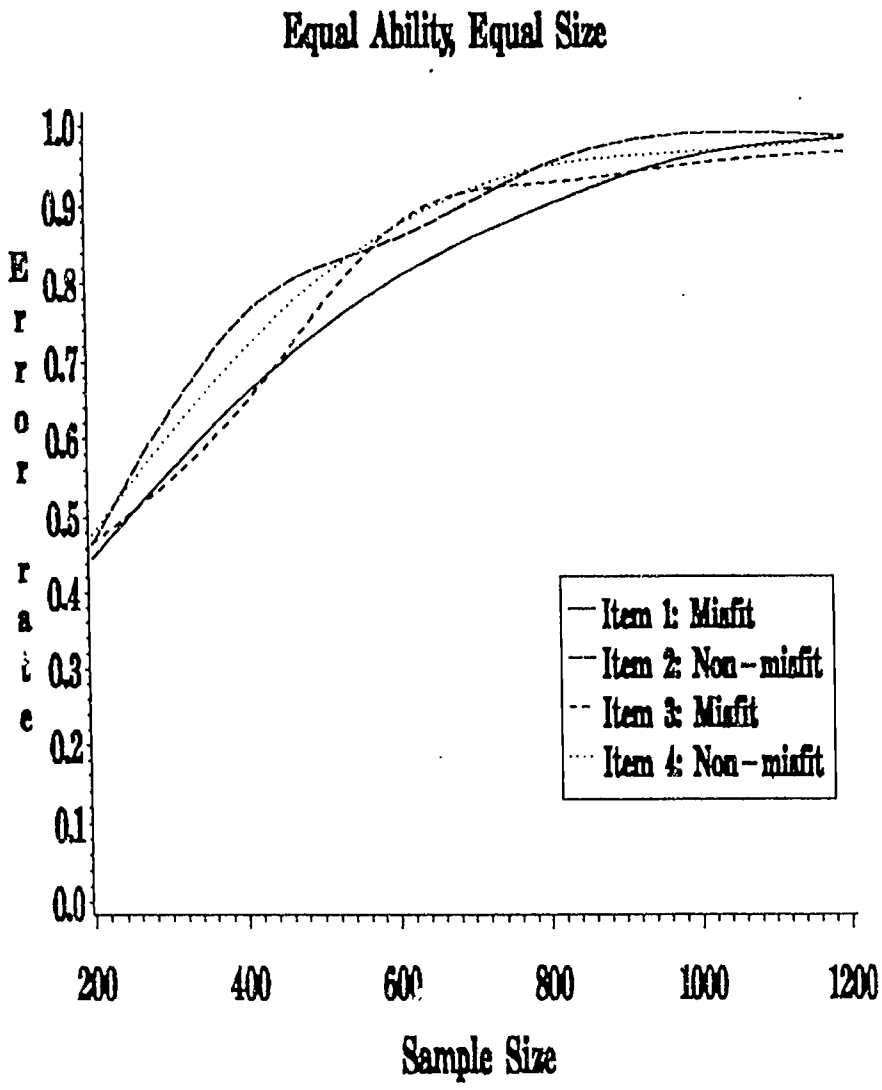
The error rate of the IRT-ANOVA method

Figure 5 displays the error rate of item 5, the unbiased misfitting item, and the mean error rate across all the unbiased, non-misfitting items (items 6 through 30). The plot on the left shows that, when there is no group ability difference, the error rate is close to its nominal level of .05. When there is a group ability difference, the error rate generally exceeds its nominal level and increases as the sample size increases. Inspection of the data shows that most of the items flagged as significant favor the low ability group. As was mentioned earlier, item 5 is simulated as a misfitting item such that the low ability group is to score higher than expected on this item based on their overall test performance, and the high ability group is to score lower than expected. The higher error rate of this item may be caused by the confounding of DIF with model-data misfit. The advantage of the low ability group and the disadvantage of the high ability group effected by a second dimension is unaccounted, or under-accounted, for by the 'unidimensional' procedure.

The plot of the error rate for non-misfitting items shows that, when there is no group ability difference, the error rate is close to its nominal level. When there is a group ability difference, however, the error rate increases slightly as the sample size increases.

Relative Efficiency of IRT-ANOVA to MH Method

In this investigation, the efficiency of the IRT-ANOVA method relative to the MH method is defined as the ratio of the rejection rate of the IRT-ANOVA method over the rejection rate of the MH nethod. A ratio greater than 1 would indicate that the IRT-ANOVA method is more powerful than the MH method, whereas a ratio less than 1 would indicate the IRT-ANOVA method is less powerful.



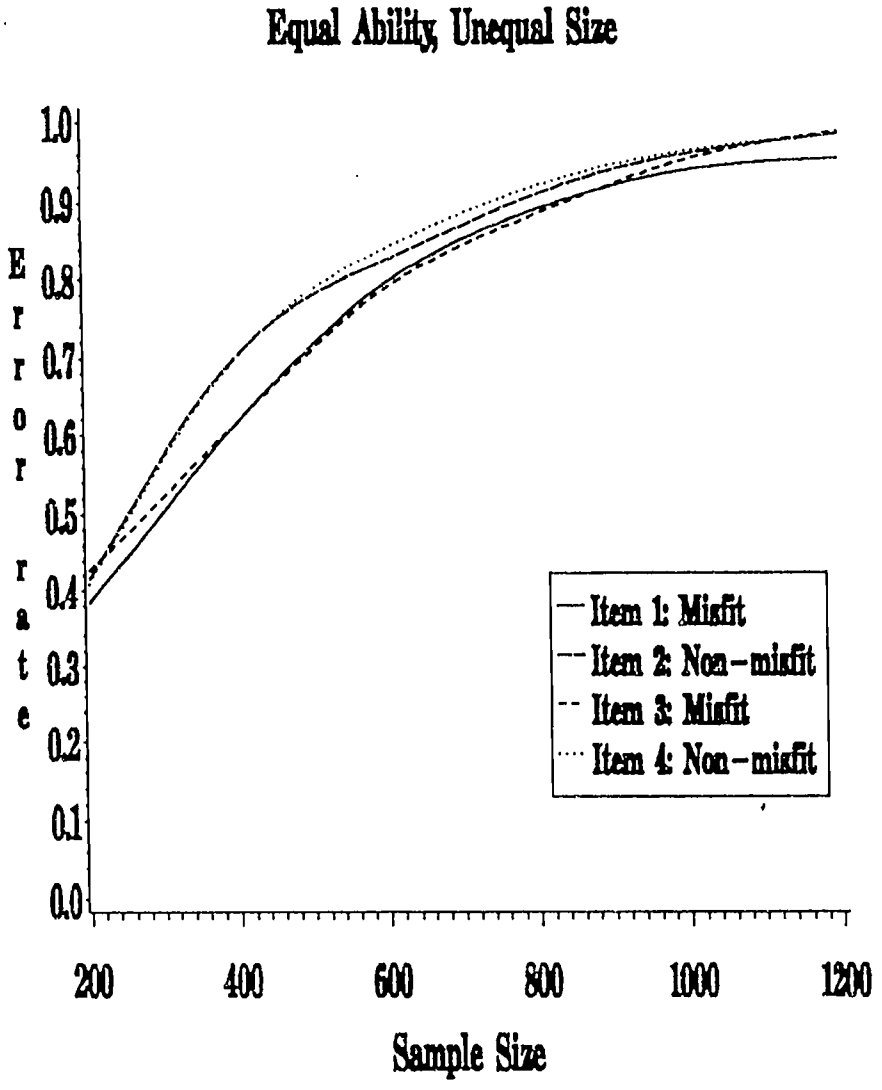


Figure 4. Rejection rate for misfit and non-misfit items when there is no group ability difference

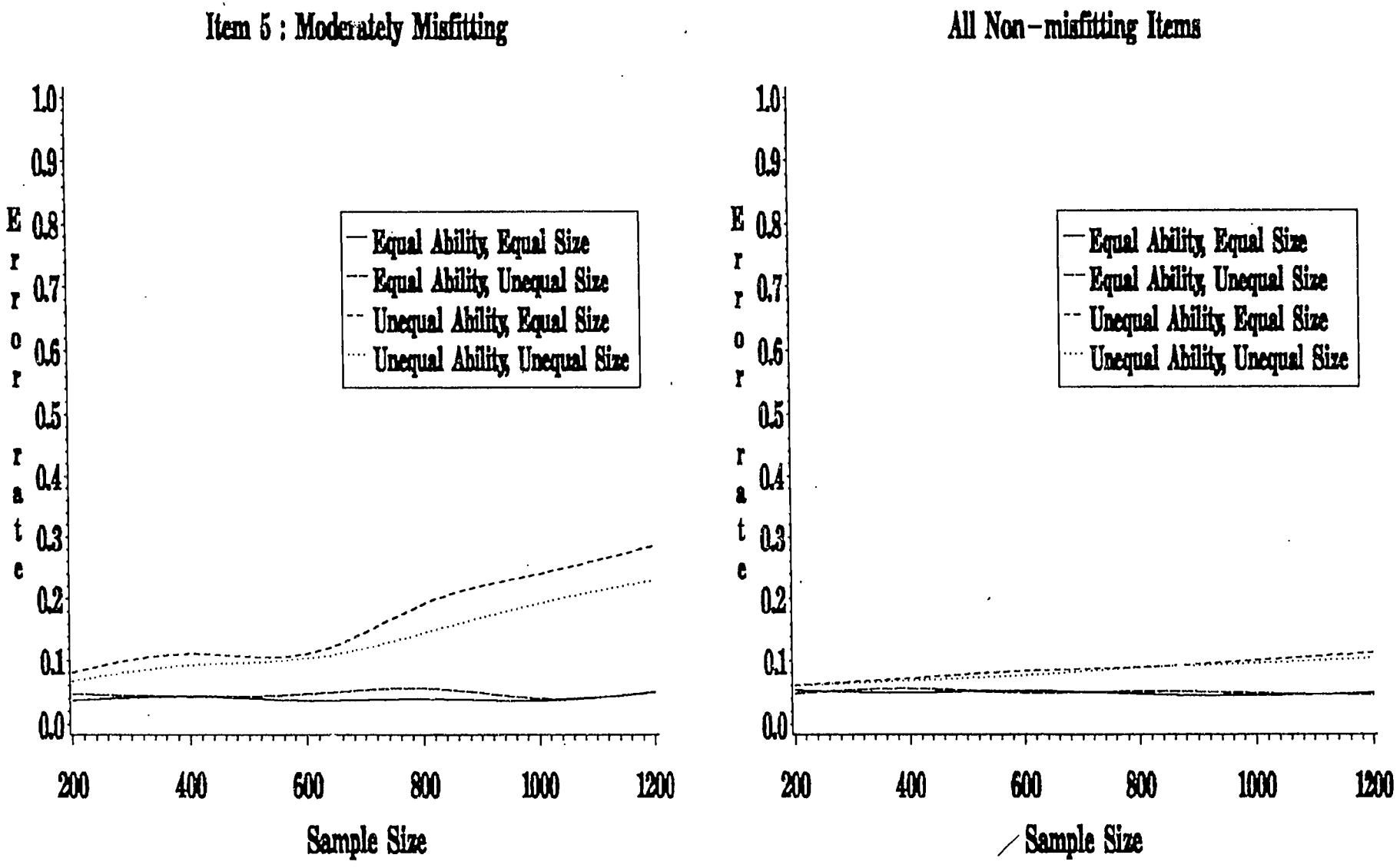


Figure 5. Error rate of the misfitting item and the mean error rate of all the non-misfitting items

Figure 6 displays the efficiency curve for each DIF item under each condition. It shows that the IRT-ANOVA method is generally more powerful than the MH method when sample sizes are small. The only exception is Item 3 under conditions when there is a group ability difference. The MH method is shown to be more powerful. Recall that for this item, the rejection rate is reduced due to lack of unidimensionality. This may imply that, while both methods assume unidimensionality, the IRT-ANOVA is more sensitive to departure from unidimensionality than the MH method.

V. SUMMARY AND CONCLUSION

This paper presents a new DIF method which employs both IRT and ANOVA. The desirable features of the procedure include: (1) concurrent calibration, (2) no ability grouping, (3) capability to examine interaction effects, (4) applicability to both dichotomous and polytomous items, and (5) applicability to relatively small sample sizes. The simulation studies show that (1) increases in group size difference decrease power; (2) DIF favoring the high ability group is easier to identify than DIF favoring the low ability group; (3) model-data misfit may increase or decrease power depending on the type of misfit and the direction of DIF; (4) the effect of misfit on power and error rate is relatively small when there is no group ability difference; (5) the error rate is higher when there is group ability difference; and (6) for small sample sizes, the method is genenerally more powerful than the MH method when data fit the model.

Research is currently under way for investigating (1) the use of the IRT-ANOVA method to examine interaction effects, (2) the effect, on both power and error rate, of group size differences in multiple-group situations, (3) the effect of item difficulty and different types of misfit, and (4) the effect of the distributional characteristics of the residuals. It is hoped that this method will contribute to DIF analysis in general, and DIF analysis in small sample situations in particular.

REFERENCES:

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, Winter 1989* (pp. 5-11).

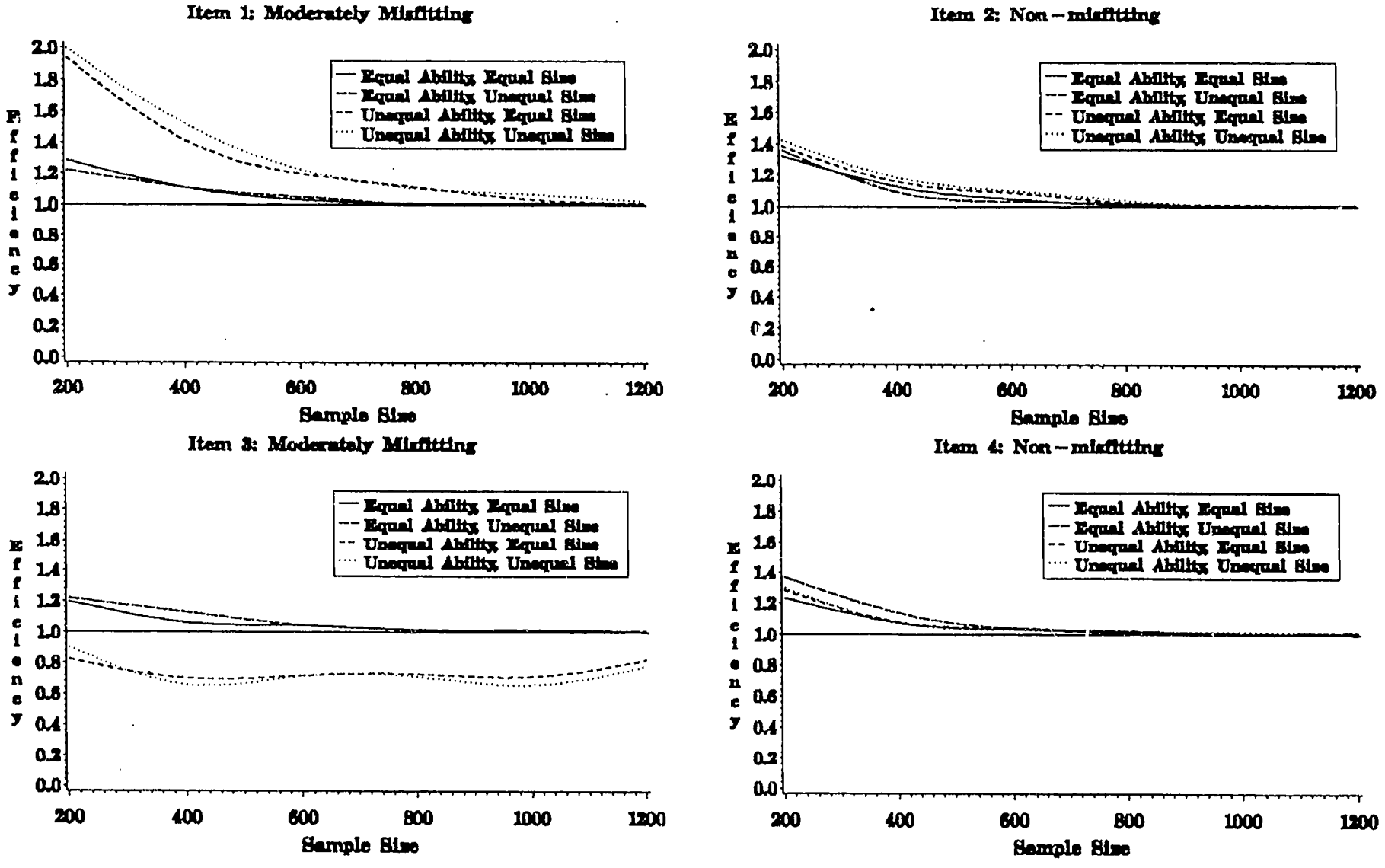


Figure 6. Relative efficiency of IRT-ANOVA to MH method