DOCUMENT RESUME

ED 356 264 TM 019 693

AUTHOR Hambleton, Ronald K.; And Others
TITLE Advances in the Detection of Differentially Functioning Test Items.
PUB DATE 93
NOTE 52p.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Educational Research; Effect Size; Guidelines; *Item Bias; Item Response Theory; *Psychometrics; *Sample Size; Statistical Significance; *Test Items; Test Validity
IDENTIFIERS *Mantel Haenszel Procedure; Power (Statistics)

ABSTRACT

The development and evaluation of methods for detecting potentially biased items or differentially functioning items (DIF) represent a critical area of research for psychometricians because of the negative impact of biased items on test validity. A summary is provided of the authors' 12 years of research at the University of Massachusetts (Amherst) pursuing item response theory-based and Mantel Haenszel (MH) DIF detection methods. In addition, a set of guidelines is offered for conducting DIF studies based on these research findings. These recommendations include the following: (1) the two-step procedure recommended by P. W. Holland and D. T. Thayer (1988) is preferred to the simple procedure; (2) the criterion used for matching examinees must be approximately unidimensional; (3) larger examinee samples are preferred; (4) with limited sample size, the power of the statistic can be increased by increasing the majority group and holding the minority group constant; (5) with very large samples, measures of both statistical significance and effect size may be important in screening items; (6) examinees in the sample must represent the population of interest; (7) combining score groups in the matching criterion may be useful for increasing the power of the MH statistic; (8) items that have lower a-parameters (item discrimination parameters) are less likely to be identified by the MH method; and (9) the MH method is not blind to non-uniform DIF. (Contains 5 tables, 3 figures and 47 references.) (SLD)

Advances in the Detection of Differentially Functioning Test Items

Ronald K. Hambleton
Brian E. Clauser
Kathleen M. Mazor
Russell W. Jones
University of Massachusetts at Amherst

Abstract

The development and evaluation of methods for detecting potentially biased items or differentially functioning items (DIF) is a critical area of research for psychometricians because of the negative impact of biased items on test validity. The purposes of this paper were (1) to provide a summary of our twelve years of DIF research at the University of Massachusetts pursuing, principally, IRT-based and Mantel-Haenszel DIF detection methods, and (2) to offer a set of guidelines for conducting DIF studies based upon our research findings.

Advances in the Detection of Differentially Functioning Test Items[1]

Ronald K. Hambleton
Brian E. Clauser
Kathleen M. Mazor
Russell W. Jones
University of Massachusetts at Amherst

Paper-and-pencil tests are widely used in selection, promotion, competency, certification, and licensure decisions throughout education, industry, and the armed services. As test use in important decision making has increased, and legal challenges to the uses of tests have become common, the question of bias in the test items has become a central concern in the assessment of test validity. Conducting an item bias study has become as common in test development and test evaluation in the United States as an item analysis or test reliability study. To this end, various judgmental and empirical methods for detecting potentially biased items have been proposed in the measurement literature (see, for example, Berk, 1982; Dorans & Holland, 1992; Hills, 1989; Mellenbergh, 1989; Scheuneman & Bleistein, 1989).

Although no statistical or judgmental method can detect "bias" as such, many statistical and judgmental methods are being used to detect items that are functioning differentially in two groups of interest (e.g. males and females). The groups are often referred to as majority and minority groups, or reference and focal groups, and the studies are referred to as studies of differential item functioning (DIF). Once a set of test items are identified as DIF, further study can be carried out to determine the most likely cause or causes of the DIF (see, for example, Scheuneman, 1987). Then, appropriate

---

[1]Laboratory of Psychometric and Evaluative Research Report No. 237. Amherst, MA: University of Massachusetts, School of Education.

action can be taken, and when necessary, the defective items can be removed from a test, or the item pool used in test development.

For the past twelve years, in our Laboratory of Psychometric and Evaluative Research at the University of Massachusetts, we have had an on-going program of research on the methodologies associated with DIF studies. The primary purposes of this paper are to present a summary of our DIF research studies and to offer a set of guidelines based upon our research for conducting DIF studies.

Our DIF research program began in 1981 with a DIF study conducted on the New Mexico State Proficiency Test, and was followed by a literature review and then the development of the "plot method". The development of the plot method which is an item response theory (IRT) based method was our attempt to by-pass the prevalent trend at the time to use statistical significance tests to flag potentially biased test items (or DIF). Our concern was that with a large enough sample size (and large sample sizes were being encouraged with the IRT-based DIF methods) even the most trivial differences between majority and minority groups would be identified as statistically significant. In effect, conscientious researchers (i.e., those who were aggressively pursuing large samples for their DIF studies) were being penalized for using large samples in their DIF studies: after "flagging" many items as DIF, they were then faced with the problem of either discarding many items or explaining why these items were being retained. Many controversies arose from these resolutions. The opposite situation was even more of a problem. Some researchers were using very small samples and, with so little statistical power, they often failed to detect sizable numbers of items showing substantial amounts of DIF.

After several years of DIF work with IRT-based methods including the plot method and the area method (both methods will be described in detail later), and the development of several computer programs, our research shifted from a comparison among IRT-based methods, to the study of DIF with the Mantel-Haenszel (MH) method. IRT DIF detection methods were rather complicated and tedious to carry out and sample sizes needed to be large. Also, there were many practical problems that arose in implementation. In contrast, the MH method had become quite popular in the late 1980s because of its simplicity, intuitive appeal, and promise. It was also being recommended by the most influential testing agency in the United States, the Educational Testing Service. Recently, two UMass doctoral dissertations and six empirical studies involving the MH method were completed. Also, a study to address the development of judgmental methods for the identification of DIF was completed. This total set of studies will be described in the remainder of the paper.

Background

In 1981 the first author was contacted by the New Mexico Department of Education about a DIF study that had been conducted for the Department on their state proficiency test. Essentially, the Rasch model had been fit to a 150 item test and the items that failed to be fit by the model were labelled as DIF (see Durovic, 1975). The problem for the New Mexico Department of Education was that 80 of the 150 items had been identified as DIF! This was a remarkable result and a result that was very hard to believe since considerable care was taken in the development of test items. A state high school proficiency test with over 50% of the items identified as DIF was a

serious problem and needed to be addressed. Were this a correct finding, very likely the test results could not have been released even after the deletion of 80 items. But the study as it was carried out had a major flaw. The study failed to distinguish between IRT model misfit and DIF. That many items were not adequately fit by the Rasch model was hardly an indication that the test was fraught with problems of bias. Our hypothesis was that many of the items labelled as DIF were items which tended to have item discrimination indices that deviated from the average level of item discrimination in the test. (Recall that in fitting the Rasch model to test data the assumption is made that all test items have similar discrimination indices.) This hypothesis was confirmed (see Figure 1) by a "U"-shaped distribution between item misfit statistics using the Rasch model and item discrimination (see Hambleton & Rogers, 1991; Hambleton & Swaminathan, 1985). We discovered that items

- -
Insert Figures 1 and 2 about here.
- -

with the highest and lowest (classical) item discrimination indices were the ones being identified as DIF.

The first thing we did was fit a two-parameter IRT model to the test data and then look at the item misfit statistics again. The findings are represented in Figure 2. The results were substantially different. This time the item misfit statistics were considerably lower (indicating better fit), and there was no relationship between item misfit and item discrimination, or item misfit and item difficulty. With this study behind us, we became interested in how best to conduct DIF studies. Our concerns were whether classical or modern measurement methods were best, and of the best methods, how they should be implemented in practice.

Our next activity was to look carefully at the DIF research literature because it seemed clear to us that DIF studies should not be equivalent to model misfit studies, and must involve the comparison of item performance data for majority and minority groups when the two groups are matched on ability. Mellenbergh (1989) and others have referred to these methods that match majority and minority groups on ability as "conditional methods". IRT-based and Mantel-Haenszel methods are conditional methods and these appear to be the methods of choice today.

A widely accepted definition of DIF is that an item is DIF (or potentially biased) if examinees of equal ability, but from different subgroups (for example, males and females) do not have an equal probability of correctly responding to that item (Hambleton & Rogers, 1989). "Differential item functioning" has become the popular term because this term focuses on the results of the analytical procedure rather than making inferences about the effect, as is the case with the term bias. If the discrepancy in item performance between the subgroups of interest is equal across the entire range of abilities then the DIF is said to be "uniform." However, if the difference between the subgroups is not consistent across the entire range of abilities then the DIF is said to be "non-uniform." Figure 3 illustrates the difference between uniform (Figure 3a) and non-uniform (Figure 3b) DIF.

- -
Insert Figure 3 about here.
- -

Figure 3a shows that on the item in question the minority group (subgroup 2) performs consistently (uniformly) lower than the majority group (subgroup 1) at all ability levels. Figure 3b shows that on the item in question the minority group performs lower than the majority group at the higher end of the ability scale and performs higher than the majority group at the lower end of

the ability scale. Thus the differences are inconsistent (or non-uniform). This distinction is important because the effectiveness of some statistical methods in the detection of DIF varies according to whether the DIF is uniform or non-uniform. It is also the failure to recognize the two types of DIF that explains some of the anomalous DIF results in the measurement literature.

Literature Review

In our review of the literature on DIF, three categories of statistical methods were identified: methods using classical test theory, methods using item response theory, and methods using chi-square methods. Readers are referred to University of Massachusetts dissertations by Clauser (1993) and Rogers (1989) for comprehensive reviews of the methods, or papers by Dorans and Holland (1992), Hills (1989), and Scheuneman and Bleistein (1989). Only a brief review of the methods in each category follows. The primary intent of this material is to set the stage for our research program.

Methods Using Classical Test Theory

A number of methods for the detection of DIF have been developed from the principles of classical test theory. These methods typically use examinee observed scores as a criterion and involve a comparison of p-values (i.e., classical item difficulty values) for the subgroups of interest. Unfortunately, classical methods are sample dependent: Draw different samples from the groups of interest and the DIF results could change. This is a potentially serious problem because the results from a DIF study using classical methods cannot be safely generalized to the larger populations of interest from which the samples are drawn.

Methods which utilize classical test theory include analysis of variance and correlational methods, the transformed item difficulty or delta plot

method (Angoff, 1982), partial correlation methods (Stricker, 1982), and the standardization method (Dorans & Kulick, 1983). Our assessment was that this general line of methods (excluding the standardization method which we liked because it was a conditional method) would not be useful to major testing programs and others agreed (see, for example, Ironson, 1983).

Methods Using Item Response Theory

Conversely, methods for DIF detection which operate within the framework of IRT not only appear to overcome the shortcomings of classical test theory, but also gain several desirable characteristics inherent within IRT. These include item statistics (specifically, item characteristic curves) which are independent of the groups from which they are derived, estimates of examinee ability which are independent of test difficulty, models which facilitate the matching of test items to examinee ability level, and models which do not necessitate the use of parallel tests for the assessment of reliability (see, for example, Hambleton, 1989; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980).

A considerable amount of research (e.g., Ironson & Subkoviak, 1979; Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Averill. 1981; Subkoviak, Mack, Ironson, & Craig, 1984) has shown (though not consistently) that methods of DIF detection which apply IRT are superior to those methods which apply classical measurement theory Essentially, these DIF detection methods involve the comparison of item characteristic curves estimated in each group of interest (see, for example, Figures 3a and 3b). An ICC represents the likelihood that examinees will get an item correct expressed as a function of their ability. If ICCs for the subgroups of interest are identical then no DIF is present for that item, however if ICCs differ between subgroups then the item is potentially biased. The parameters which contribute to the ICC

can include: item difficulty, or the b parameter, represented on the ICC plot as that position on the ability scale corresponding to the point of inflection (the point at which the ICC changes from curving upward to curving downward); item discrimination, or the a parameter, represented on the ICC as a value which is proportional to the slope of the ICC at the point of inflection; and the pseudo-guessing parameter, or the c parameter, which is the probability of answering an item correctly regardless of how low the ability, and is represented on the ICC as the lower asymptote of the ICC.

The theoretical underpinnings of IRT led it to become one of the most promising methods of DIF detection (Ironson, 1983) to emerge during the 1980s. The main advantage is that item performance differences between subgroups of interest can be studied, conditional on ability level.

Several different methods for determining the similarity of ICCs have been developed, however, as Scheuneman and Bleistein (1989) conclude in their comprehensive review of DIF detection methods, "no clear preference for any one of these [methods] has emerged" (p. 267). These methods include the difficulty shift method (Wright, Mead, & Draba, 1976) and the analysis of fit method (Durovic, 1975) which are based on the one-parameter, or Rasch, model and the IRT area method (Ironson & Subkoviak, 1979) and the two stage method (Lord, 1980), which are based on the three-parameter model. These methods involve a comparison of ICCs in the subgroups of interest using the item parameter estimates, or areas between the ICCs or the "sum of squared differences between the ICCs" over the portion of the ability scale of interest. An additional variation among the methods concerns the use of weights to reflect the minority and/or the majority ability distributions. Another IRT-based DIF detection method is the plot method (Hambleton & Rogers, 1991).

Methods Involving Chi-square Analyses

A number of DIF detection methods make use of a chi-square value as an index of DIF. These procedures are often collectively referred to as chi-square methods. Chi square methods include the modified contingency table method (Veale, 1977), the Mantel-Haenszel method (Holland & Thayer, 1986; 1988) which, due to its widespread adoption, will be discussed in greater detail later in this paper, the logistic regression method (Swaminathan & Rogers, 1990), and the contingency table method (Scheuneman, 1979). With all of these methods, the test score, or "purified test score" (with potentially biased test items eliminated) is used to match the majority and minority groups prior to comparing item performances.

IRT Area Method

With our active research program at UMass in the 1980s on IRT models and their applications (see, for example, Hambleton & Swaminathan, 1985) it was logical for us to look at the applications of IRT to the study of DIF detection. The research was prompted by two problems which we identified in the DIF literature at the time. First, while the use of item response models appeared to hold considerable potential for detecting DIF, many researchers were using the less defensible one-parameter model. The one-parameter model is highly restrictive and so model-data fit may be confounded with the results of DIF studies using the one-parameter (or Rasch) model. This was the finding from our analysis of the New Mexico data described earlier in the paper. Also, when the model-data fit is poor, comparisons of item difficulty values in majority and minority groups will not be very useful because the item difficulty values themselves provide inaccurate information about the functioning of test items in the two groups. A second problem arises when

significance tests are used in DIF studies. The results of significance tests are influenced by examinee sample sizes. With an examinee sample size of 100, perhaps no truly biased items will be identified; with a sample size of 5000, it is possible that all items will be identified as biased despite the fact that the actual item performance differences may be of no practical consequence (Hambleton, 1989). The plot method appeared to overcome both problems and therefore it seemed worthy of full development and study.

The origin of the IRT plot method is unknown to us, though the general approach using non-IRT concepts was described by Angoff (1982). Perhaps, though, Shepard (1981) was the first to introduce the general approach into the IRT literature. Basically, the researcher first finds an IRT model that fits the test data. Then the researcher generates a plot of item difficulty values for randomly equivalent groups and then compares the plot to the plot of item difficulty values for two other groups, i.e., the majority and minority subgroups. The first plot provides a baseline for interpreting the second plot. If the two plots are similar, then it is reasonable to assume that the subgroups of interest are no more different in their response processes than the randomly equivalent groups. Subgroup membership can then be ruled out as a factor in item performance. If, on the other hand, the plots are quite different, the feature of item parameter invariance over subgroups is not present and attention must shift to identifying those test items that functioned consistently differently in the subgroups of interest. Examples appear in Hambleton and Rogers (1991) and Hambleton (1989). The method was applied successfully by us in several DIF studies including our work with the (United States) National Assessment of Educational Assessment in the area of mathematics.

Probably the most popular of the IRT-based DIF methods is the "area method." Here, the ICCs for the majority and minority groups are compared and the area between the ICCs over some interval on the ability scale is used as an indicator of DIF. The bigger the area, the more DIF is said to be present. Rogers and Hambleton (1985) developed a computer program to compute the area between two ICCs while attending to the technical problems of choosing the interval, rescaling the majority and minority group item parameter estimates to a common scale, handling "outlier" item parameter estimates, and introducing variations on the area method (1) to include a weighting option (researchers could weight the majority and minority group item performance differences at specific ability levels by the minority distribution) and (2) to handle the presence of non-uniform DIF.

One shortcoming of the area method (besides the complexities in applying the method and the need for fairly large samples) was the lack of a critical value for interpreting the area statistics. Recently Raju (1988, 1990) has successfully addressed this shortcoming but, in our work in the middle 1980s, the only option available was to rank order items based on their area statistics and then to take an especially hard look at items with the highest area statistics. Our solution at the time was to use simulation procedures to generate an area statistic sampling distribution under the hypothesis of "no DIF." This distribution could then be used to set the .05 and .01 significance levels. In our simulations, we were able to match the ability distributions, test length, sample sizes, and item parameter estimates to the real data being studied. Then we were able to simulate item performance and produce area statistics under the null hypothesis that there was no DIF present in the data. This work was described in a study by Rogers and

Hambleton (1989) and served as the basis for setting critical values for interpreting area statistics from DIF studies.

Just about the time we were feeling quite confident that we had two promising methods under control, the plot method and the IRT-based area method, the Mantel-Haenszel (MH) method was introduced into the measurement literature by ETS (Holland & Thayer, 1986, 1988). This led us to design a study to compare the MH method with the IRT-based area method, since the latter was the preferred method of those using IRT methods.

The main purpose of the study was to carry out a detailed analysis of the results of applying the IRT-based area method and the MH method to the same set of test data. Our interest centered on the degree of agreement between the methods in identifying DIF and the possible reasons for disagreements when they were found. The research was primarily intended to determine the consequences of substituting the easier-to-use and more convenient MH method for the "theoretically preferred" but difficult to implement in practice IRT-based area method.

The research reported in Hambleton and Rogers (1989) was carried out using two samples of 1000 students of Native American and Anglo-American students who took the 1982 version of the New Mexico State Proficiency Test. Both DIF methods were implemented using the very best techniques possible. For example, large samples were used to increase the stability of DIF results (which is often a primary reason for disagreement among DIF methods), and only items which were identified as DIF in two parallel DIF studies with the same method were considered in the study of agreement across methods.

- -

Insert Table 1 about here.

- -

Table 1 contains a summary of the main comparison in the paper. Of the 16 test items consistently identified as DIF by one method or the other, seven test items were in agreement (i.e., identified consistently by both methods as DIF). Of the remaining nine, three items were nearly in agreement but were not, probably because of type II error. And here is the surprising result: Of the remaining six test items, four items were identified by the IRT area method as showing non-uniform DIF. These four test items were not detected by the MH method. The main finding of the study seemed to be that the two methods showed high agreement in the detection of DIF items. The disagreements were due mainly to the inability of the MH method to detect non-uniform DIF (a weakness which is well known). This important finding, and in view of the ease with which the MH method could be implemented, caused us to refocus our research and begin a series of methodologically-oriented studies with the MH method.

MH DIF Detection Method

While the debate as to which statistical DIF detection method is most adequate may go unresolved, it is clear that the MH method has emerged as one of the preferred. The best answer to the question of which method should be used may be that in high stakes (i.e., important) testing situations the most adequate choice is the routine use of multiple methods. In such cases, the threat of increased type I error (i.e., labelling items as DIF when they are not) may be more easily tolerated than unnecessary type II error (i.e., failing to identify items which are DIF). Regardless of the decisions made by individual test developers, it is clear that the MH statistic will have an important place in item screening. The question then shifts from whether to use the MH statistic to how it might be most appropriately used. This section will review six studies which we carried out over the last three years to

examine aspects of that question. All of the studies used a computer program to compute the MH statistic which was developed by Rogers and Hambleton (in press). Details of the MH method are found in the papers by Holland and Thayer (1986, 1988) and will not be presented here.

Internal versus External Criterion

As was described above, the MH method allows for a comparison of the chances for success on an item, by members of two groups, after first conditioning (matching examinees) on the ability of interest. In common use, this matching is based on the total test score. This practice requires a kind of "boot strap" logic in which the validity of a set of test items is judged by a criterion based on items of undemonstrated validity. It can therefore be argued that it would be preferable to match on a previously validated external criterion. Unfortunately, such a criterion is rarely available. This raises the question of to what extent reliance on an internal criterion will produce results which differ from those of an appropriate external criterion. The first paper examined that issue.

Hambleton, Bollwark, and Rogers (in press) compared the results of a sex DIF study on a high school scholarship test using total score on the test itself (internal criterion) and scores on a high school achievement test (external criterion). The main finding, which was replicated on three additional scholarship tests in other subject areas, was that the internal and external criterion measures produced highly similar MH DIF results. In view of the moderate correlations between the internal and external criterion measures (they ranged from .38 to .52 across the four subject areas), this finding seemed to support the continued use of the internal criterion in MH studies. Certainly the agreement between methods would increase even more if the correlation between the criterion measures was higher.

Selection of the Internal Criterion

Having made a decision to use an internal criterion, issues still remain as to how that criterion should be formed. Not all items in the test need to be included in the criterion measure. Item bias may be conceptualized as a consequence of multidimensionality. If a set of test items is sensitive to more than one ability, and if there are between-group differences in the underlying ability distributions, then multidimensionality may manifest as DIF (see, for example, Ackerman, 1992). One solution to this problem which has been proposed is that homogeneous or valid subtests from the total test be selected, and that the subtest score be used as the matching criterion. It is argued that this provides a "cleaner" and more valid criterion for matching majority and minority groups. These subtests may be identified either statistically, or by judgmental review. In the study by Clauser, Mazor, and Hambleton (1991a), judgmental review was used to sort items into subtests, and then each item was analyzed within the context of both the total test, and the subtest(s) to which it was assigned.

The MH method was used to evaluate the responses of samples of 1000 Anglo-American and 1000 Native American examinees to 91 items taken from a high school proficiency test. The test contained items which required a variety of skills for a correct response. These skills included (1) reading, (2) mathematical calculation, (3) interpretation of tables, charts, or maps, and (4) certain types of prior knowledge. The first matching criterion used for the MH method was the total test score. Items from each of these four categories were then re-analyzed using only items within that category in the matching criterion. Because these re-analyses required a reduction in the number of items used in the matching criterion, three additional subtests were formed and re-analyzed. These were randomly formed subtests that differed

from the total test only by length. These were intended to act as controls to demonstrate that any effects observed in the analysis of the content-based subtests were not the result of test length alone. Finally, one additional evaluation was conducted. This was based on a subgroup of items from the "prior knowledge" category which were poorly written, resulting in the case that no clearly best answer was available. It was assumed that these items required guessing as to what the item writers intended. Guessing in this context was considered to be a nuisance ability.

- -
Insert Table 2 about here.
- -

Table 2 summarizes the results of these analyses. These results suggested the presence of two apparently conflicting effects. When items identified as displaying DIF on the full test were re-analyzed, as part of content-based subtests, 32% ceased to be so identified. This suggests that to avoid type I error, it may be prudent for test developers to screen items with the MH method in the context of content-based subtests. It was additionally noted that, across both content-based and randomly formed subtests, more items were identified when analyzed as part of subtests than as part of the full test. It would appear that the critical variable in this context is test length. More research is needed to fully explain this phenomenon. One final result worth considering from the study was the performance of those "prior knowledge" items judged to have "no clearly best answer." Of the items in this category, 50% were identified as displaying DIF. When these items were re-analyzed as part of the "prior knowledge" subtest, they all continued to be identified as displaying DIF. This suggests that careful attention to the matching criteria may help practitioners to distinguish between problematic items and those for which DIF identification represents a kind of artifact.

Power and Sample Size

The remaining four studies examined variables which related to the power of the MH statistic to identify DIF. All four were based on simulated data. This was necessary to allow for a distinction between increased type I error and increased power.

The first of these studies examined the sample size variable (Mazor, Clauser, & Hambleton, 1992). The data were generated using a three-parameter logistic IRT model. To create conditions that were representative of those found in practice, a test was produced containing 59 items generated with a and b parameters based on estimated values from a recent administration of the GMAT (Kingston, Leary, & Wightman, 1988). C-parameters for all items were set at 0.20. Sixteen additional studied items were added to the 59, to create a total test of 75 items. These studied items were created to represent a wide range of item types. Four a-parameter values (.25, .60, .90, 1.25) were crossed with five reference group b-parameter values (-2.5, -1.0, 0, 1.0, 2.5). Uniform DIF was modelled with four levels of difference in the b-parameter value between groups (.25, .50, 1.00, 1.50). Crossing these conditions produced a total of 80 studied items. Combining these items, 16 at a time, in the test with 59 non-studied items produced five 75-item tests.

Examinee responses for these sets of item parameters were simulated for sample sizes of 100, 200, 500, 1000, and 2000 per group. This was done first for majority and minority groups with abilities that were normally distributed with equal mean. (This is similar to conditions typically found in male-female comparisons.) The simulations were then repeated with ability distributions of the same shape but with the minority group mean one standard deviation below that of the majority group. This is within the range typically reported for comparisons involving various ethnic groups (Raju,

Bode, & Larsen, 1989; Hambleton & Rogers, 1989). The MH statistic was then used to evaluate each of the data sets. The procedure used was the two-step process described by Holland and Thayer (1988).

In one sense, the results of this study were not surprising. Increased sample size was associated with increased power. However, these results provided practical insights into the power of the statistic across various conditions which were considerably less trivial. The MH method has been described as the method of choice for use with small samples. Hills (1989) recommended its use with samples as small as 100. The results of the simulations carried out in this study suggested that such a recommendation could not be justified.

To provide a practical metric for judging the impact that DIF of the size simulated might have in a test situation, p-values were calculated for majority and minority groups of equal ability. Table 3 shows the largest p-value difference between groups missed as well as the smallest difference

- -
Insert Table 3 about here.
- -

identified under each sample size condition. This table also shows the percentage of items identified under each condition. With sample sizes as small as 100 examinees per group, items were missed with p-value differences between groups in excess of .20. In many testing situations, this difference would not be considered acceptable. Ten items of this type on a test could lead to a difference of two points on the total test. Practitioners can gain insight into the actual sample size required for appropriate screening in their testing situation from Table 3. Clearly, there is no single correct answer. Larger samples will result in more sensitive screening. Our view was

that samples between 200 and 1000 per group would be sufficient for most purposes.

Power and Item Characteristics

The next study re-examined this same data set, asking the question, "are there certain types of items that display DIF, but are likely to be missed by the MH statistic?" While the previous study provided a good indication of the size of item performance differences that might be detected or not detected as a function of sample size, our next concern was the statistical characteristics of DIF items that might go undetected. The study by Clauser, Mazor, and Hambleton (1991b) focused on the results for samples of 1000 examinees per group. The results indicated that not only the difference between the b-parameters (item difficulty parameters) for the two groups, but the absolute value of the b-parameter and the value of the a-parameter influenced the likelihood that an item would be identified as DIF. Lower a-parameter values (item discrimination parameters) were associated with items that were likely to be missed regardless of the difference in the b-parameter values between groups. Similarly, very difficult items were likely to go undetected, regardless of the level of DIF modelled. This was particularly the case with unequal ability distribution comparisons. This apparently resulted from the fact that, under such conditions, there are too few examinees at the part of the ability scale where such items are functioning. In spite of this discrepancy across ability distribution conditions, we noted that, although the MH method was most effective with examinees from groups of equal ability, it remained useful with groups of considerably different ability (note the results in Table 3). For the practitioner, these results suggest that the MH method will be effective over a wide range of conditions. However, removal of items from a test based on their high MH statistics would

result in the disproportionate removal of the most discriminating items. In spite of this loss in discrimination in the test, considerable undetected DIF could remain in items with similar characteristics but lower item discrimination. Additionally, practitioners should be concerned with the effect noted for particularly difficult items. In cases where a test is intended to identify a small number of extremely competent examinees, such items could be a major problem. DIF screening should be carried out using a sample from the population of interest. If that population is extremely competent examinees, that sample should be used. Screening items using a less competent sample could result in missing DIF items in the more capable examinee populations.

Width of Score Groups in Matching Criterion

A follow-up study was also carried out by Clauser, Mazor, and Hambleton (in press) using this same simulated data set. This study examined the potential for increasing the power of the MH statistic by reducing the number of score groups used in the matching criterion. Holland and Thayer (1988) recommended that k+1 score groups be used, where k is the number of items on the test. It was subsequently suggested that the power of the method might be increased if the number of score groups used in matching was reduced. As few as four or five score groups have been recommended as optimal with other chi-square type DIF detection methods. To examine this variable, the data set described above was re-examined first using k+1 (76) score groups in the matching criterion. This analysis was then repeated using 20, 10, 5, and 2 score groups. Because it was of interest to measure changes in the type I error rate, an additional data set was simulated. This had a-parameter (.60, 1.00, 1.25) and b-parameter (-2.50, -1.00, 0, 1.00, 2.50) values similar to

those in the DIF items, but no difference in b-parameter values between groups.

The results suggested that modest gains in power could be produced by reducing the number of score groups in the matching criterion, if the majority and minority groups had similar ability distributions. A reduction from 76 to 5 score groups resulted in a 2% increase in the DIF identification rate, when a sample size of 1000 per group was used. Further reduction to two score groups resulted in an additional 3% increase. Similar increases in identification rate were evident at other sample sizes. For example, with 200 examinees per group, a reduction from 76 to 2 score groups resulted in a 3% increase in DIF identification. While these increases in power are modest, they appeared to be associated with little or no increase in the type I error rate.

By contrast, substantially larger increases in the identification rate were observed with the unequal ability distribution comparisons. Again, for a sample size of 1000, a reduction from 76 to 5 score groups resulted in a 10% increase. An additional 7% gain was noted when further reducing to 2 score groups. With a sample size of 200, an 8% gain was associated with the reduction from 76 to 5 score groups. This increased to 18% when reducing from 76 to 2 score groups.

Unfortunately, these gains did not occur without inflated type I error. When the full number of score groups was used, no type I error was noted. With a reduction to 5 score groups and a sample size of 1000, the type I error rate was 33%. Using two score groups resulted in a 67% type I error rate. Examination of the data suggested that this increased type I error resulted from a contamination of the matching criterion under these conditions. The MH method makes comparisons based on the assumption that examinees from the two

groups, within any score category, have equal observed scores. With k+1 score groups, this assumption is met, regardless of the ability distributions for the two groups. If examinees are being compared from groups with equal ability distributions, reduction in the number of score groups introduces only a minimal violation of this assumption. When examinee groups of different ability are compared, this assumption will be increasingly violated as the width of the score groups increases.

These results suggested that more than a modest reduction in the number of score groups used in matching cannot be recommended when comparisons are being made between examinee groups of dissimilar ability. Increasing sample size is a preferred means of increasing the statistic's power. Practitioners choosing to use fewer than the maximum number of score groups should be aware of the threat to the validity of the matching criterion that can result when dissimilar examinee groups are compared.

Non-Uniform DIF Detection

One repeated criticism of the MH statistic is that it is not useful for detecting non-uniform DIF (Hills, 1989; Swaminathan & Rogers, 1990). The last of the six studies (Mazor, Clauser, & Hambleton, in press) presented in this section focused on the use of a variation on the MH procedure that is sensitive to this type of DIF. In view of the generally positive results obtained with the MH method, our goal was to attempt to overcome the one main shortcoming - the failure of the MH method to identify non-uniform DIF.

Non-uniform DIF results from an interaction between ability and group membership. In the context of a three-parameter IRT model, this type of DIF would produce ICCs for the two groups which cross (see Figure 3b). Because the MH method is a signed statistic, positive differences in one part of the ability distribution can offset negative ones in another. The variation of

the MH method examined in this study was a three-part procedure. In the first step, the MH method is run (as usual) on the full majority and minority group sample. The sample is then split based on the combined mean for the two groups. The MH method is then run on all examinees with a score equal to or less than the mean. It is finally repeated for examinees with scores above the mean. To examine the utility of this modified MH method, the simulated data set described above had to be expanded to allow for differences in the a-parameter between groups. The new data set included four levels of the a-parameter (.25, .60, .90, 1.25), five levels of the b-parameter (-2.5, -1.0, 0, 1.0, 2.5), five levels of difference in the a-parameter between groups (0, .25, .50, .75, 1.0), and four levels of between-group difference in the b-parameter (0, .3, .6. 1.0). Crossing these conditions produced 400 studied items. As with the previous simulations, these items were combined, 16 at a time, with 59 non-DIF items to form a series of 75-item tests. The sample size was 1000 examinees per group.

The results confirmed the findings of previous studies (Rogers, 1989) that the typical MH method was able to identify a substantial number of non-uniform DIF items with ICCs that crossed away from the center of the ability distribution, but essentially none of the items with ICCs crossing at the middle. Table 4 shows that the modified (i.e., three-step) MH method was able to identify a substantial number of the previously undetected DIF items. The results also suggested that these advantages were gained without a discernible inflation in type I error. While this was a exploratory study, and additional research is necessary, the findings appeared to be very promising.

- -
Insert Tables 4 and 5 about here.
- -

Summary

The six studies described in this section are summarized in Table 5. The various results support the continued use of the MH statistic for DIF identification. The first two studies were focused on the matching criterion. They suggest that using an internal criterion for matching examinees on ability is appropriate. However, the validity of this approach depends on the adequacy of the total test score as a measure of the ability of interest. When individual items measure more than one ability, or when items measuring different abilities are part of one test, the adequacy of the matching criterion may be compromised, leading to errors in identification of DIF.

The four simulation studies focused on the power of the statistic under various conditions. The first study confirmed the usefulness of the MH method with relatively small samples, but it did not support its use with samples below 200 per group, except in cases where only the roughest measure of DIF is required. The second study highlighted that the power of the MH method is related not only to the difference in item difficulty between groups, but to the item's discrimination and to the interaction between item difficulty and the ability distributions of the examinee samples.

The third simulation study examined the utility of using a reduced number of score groups in the matching criterion. The results suggested that some increase in power may result, but careful attention must be paid to the examinee ability distributions. The validity of the matching criterion may be damaged when score groups are collapsed for examinees from groups with significantly differing abilities. The final study suggested a variation on the MH method which allows for identification of non-uniform DIF. The initial results were promising, although more research is needed.

Judgmental Methods

One of the most potentially useful methods for detecting the possibility of DIF is to judge each item subjectively for the presence or absence of any characteristic or feature which may lead to the item exhibiting DIF. Such methods are called "judgmental methods." Essentially, judgmental methods to DIF detection require a number of judges to be tasked with reviewing a series of items comprising a test. The judges may act independently; however, frequently groups of judges are brought together for useful and productive discussion. In addition to issues of stereotyping and fair representation, the focus of the judges' attention should be directed towards ensuring that examinees belonging to both majority and minority groups have equal familiarity and experience with item content (Tittle, 1982). In this way judgmental approaches to DIF detection perform the role of establishing construct validity (Tittle, 1982), content validity, and face validity. All valuable and important aspects of test development.

Typically, judgmental and empirical methods for detecting differentially functioning items have shown little agreement. A partial explanation for this low agreement is that the judgmental review forms are sometimes focused on cultural and sexual stereotyping in items rather than on factors which may lead to differential performance between subgroups of interest. As a result, many undesirable items are identified in the item bias review process, such as those which may show members of minority groups doing unskilled work or having problems of one kind or another. However, the items identified, although undesirable, are unlikely to function differentially in actual practice.

In 1988, we conducted a literature review (see Hambleton & Rogers, 1988) to identify aspects of items that ought to be thoroughly investigated during a judgmental review. These aspects, organized into two broad categories, are given below:

Stereotyping and Inadequate Representation

Does the item:

1. contain material which is inflammatory, controversial, or emotionally charged for members of minority groups?

2. contain language or material which is demeaning or offensive to members of minority groups?

3. portray members of minority groups in situations that do not involve authority or leadership?

4. depict members of minority groups as experiencing stereotyped emotions?

5. depict members of minority groups as having stereotyped characteristics?

6. depict members of minority groups in stereotyped occupations?

7. contain biased or offensive art work?

Sex, Ethnic, Cultural, Religious, and Class Bias

Does the item:

8. contain content that is different or unfamiliar to some minority groups?

9. measure what is taught in the curriculum? (for achievement tests)

10. reflect information and/or skills that may not be expected to be within the educational background of all examinees?

11. contain information that could benefit examinees of the majority group?

12. contain words which have different or unfamiliar meanings for minority groups?

13. contain group-specific language, vocabulary, or reference pronouns?

14. contain distractors which may be especially attractive to members of the minority groups for cultural reasons?

15. because of the format or structure of the item present greater problems for students from some backgrounds than from others?

With the background to design a potentially more valid judgmental review form after completing our review, a new form was designed and used in a recent comparative study of judgmental and empirical methods (Hambleton & Jones, 1993).

The purpose of the Hambleton-Jones study was to refine, in relation to common practices, both statistical and judgmental methods for detecting potentially biased items in an attempt to improve the agreement between the results obtained with these methods. This seemed a worthy goal because, if greater agreement between methods can be achieved, test items can be more effectively screened using judgmental methods prior to field testing or actual test administrations. In fact, in some small scale test development studies, item bias reviews may be as much bias identification work as can be accomplished. In other studies, empirical work can be done but the results are unstable because of small sample sizes, especially for the minority group. Also, the fewer items that are defective during field tests or test administrations, the more creditable the agencies producing the tests are judged to be. Clearly, therefore, research that might lead to improvements in item bias review forms seemed desirable.

The statistical methods in the study were refined by (1) focusing only on items which were differentially functioning in both the original sample and in a cross-validation sample, (2) carefully choosing the interval over which DIF was measured and the cut-off score for interpreting the DIF statistics, and (3) using more than one DIF statistic in the empirical analysis. The judgmental methods were refined by (1) carefully distinguishing between stereotyping of groups and factors which could differentially impact on test performance, and (2) using the findings of an earlier study by the authors to refine the item bias review form.

Several major points emerged from the analyses. First, both the IRT-based method and the MH method were somewhat unreliable in identifying differentially functioning items. This result helps to explain the moderate agreement reported in the measurement literature among approaches to DIF. The

fact is that studies of overlap of results with methods for investigating DIF are influenced considerably by the unreliability of the methods. There appeared to be substantial agreement between the IRT-based method (the IRT area method) and the MH method in the detection of DIF when only items which showed DIF in a cross-validation sample were considered in the analysis.

Second, our work appeared somewhat successful with the item bias reviews. Five of eleven items identified by the judges as potentially biased were identified as DIF by the empirical methods. With a couple of changes in the item bias review form, the agreement would have been even higher -- for example, ask judges to identify test items with negative words or ideas in the stem, and search for test items that require prior knowledge that may be less present in the minority group than in the majority group. Of course, the generalizability of these recommendations to other editions of the test we worked with or to other basic skills tests or to other ethnic groups is unknown.

The implications of the results of this study for practice seemed clear. First, test developers should be reminded about the unreliability of DIF statistics. This means that they should be encouraged to use large samples in their analyses whenever possible and interpret the statistics with a fair degree of caution. Second, the evidence suggested that the MH method can be safely substituted for IRT-based methods if safeguards are put in place to detect non-uniform DIF. Some of these items are likely to go undetected by the MH method. Finally, and most importantly, there was some evidence that a judgmental process can be effective in identifying test items that may be DIF in practice. And, careful analysis of items which are identified as DIF using empirical methods may be helpful in redesigning item bias review forms. By so doing, more effective item bias reviews can be carried out. This suggestion

seems especially applicable within an on-going testing program. How useful a "tailored" review form for one test will be for another, or even how useful the form will be for identifying multiple types of DIF remains to be determined.

Some Guidelines for Conducting DIF Studies

After 12 years of DIF detection research, our confidence in being able to properly design and conduct valid DIF studies is relatively high. Our first guideline is that there is no single method that can be guaranteed to identify all of the DIF items in a test. For that matter, neither can several methods, but multiple methods can address the instability problem which undermines the utility of current methods and can address the shortcomings found in particular methods. Careful test developers working on important tests will almost certainly want to apply several DIF methods, including a judgmental method and one or more empirical methods. Professional judgment based upon careful analyses of the results will be needed to sort out reasons for disagreements among the methods in the detection of DIF.

Our second guideline is that judgmental methods are invaluable in the DIF detection process. In fact, judgmental methods of DIF detection have several advantages over other methods. Firstly, the use of judgmental methods to DIF detection are often considerably cheaper than statistical methods which require the collection of test data. Secondly, the use of appropriate and suitable judges to review test items provides the test developer with credible face validity. In the social, ethical, and politically conscious arena in which test developers operate, the opportunity to have judges from minority groups of interest review and judge items as DIF or non DIF provides the test developer with face validity that is both valuable and powerful. Thirdly,

judgmental reviews may be performed prior to any examinees being administered the test items. Poor items can then be weeded out or modified before the items are field tested. This is particularly valuable for those test developers who are unable to field test their items. Fourthly, if judges are selected who are familiar with curricular content, then judges can also be tasked with ensuring that the test as a whole exhibits content validity. That is, the test reflects the content of the curriculum that it is designed to measure.

A number of disadvantages are inherent with the judgmental approach to DIF detection. Most notable of these is the frequent failure of statistical and judgmental approaches to agree on which items are flagged as exhibiting DIF. Other disadvantages include: the expense incurred in bringing judges together; the time and expense involved in training judges; and the susceptibility of judges to fatigue, boredom, and other conditions which may interfere with the validity and reliability of judgments. In the presence of both advantages and disadvantages, judgmental methods have a definite role to play, but they are not sufficient.

It would appear that the MH method is rapidly achieving the status of "industry standard" for DIF identification. Hills (1989) cites the following advantages for the method: It provides both a measure of effect size and a test of statistical significance; it is the uniformly most powerful and unbiased test of the null hypothesis against the alternative hypothesis that the probability of a correct response is not uniform across ability levels; it can be used with relatively small samples and with unmatched samples; and it is inexpensive to use. Additionally, it has been empirically shown to produce results which are similar to those of the IRT area method (Hambleton & Rogers, 1989). The primary disadvantage of the MH statistic is its demonstrated

inability to identify non-uniform DIF (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990).

Our third guideline, therefore, is that for most test development projects, the MH method should be used and will do more than an adequate job, especially if some of our findings below (see Clauser, 1993) are considered in the implementation:

1. The two-step procedure recommended by Holland and Thayer (1988) is preferred to the simple procedure. Purifying the matching criterion by removing items identified as displaying DIF on an initial implementation of the statistic has theoretical appeal and empirical support (Clauser, 1983).
2. The criterion used for matching examinees must be approximately unidimensional. Both Ackerman (1982) and Clauser, Mazor, and Hambleton (1991a) have shown that substantial type I error may result from violations of this assumption. If this assumption is in question for the test as a whole, the test may be broken down based on item content. MH analysis may then be carried out on approximately unidimensional subtests.
3. Larger examinee samples are to be preferred to smaller samples. Samples of less than 200 per group (majority and minority) may be insufficient for many purposes.
4. When the sample size is limited because there are relatively few minority group members, the power of the statistic can be increased by increasing the majority group while holding the minority group constant. Ratios of as much as nine to one were useful in increasing power and did not appear to be associated with increased error or other bias (see Clauser, 1993).

5. When very large samples are used, it may be important to use measures of both statistical significance and effect size in screening items. Because the power of the statistic increases with sample size, with samples in excess of 1,000 per group, trivial levels of DIF may be identified as statistically significant. In such cases, a system such as that currently in use at Educational Testing Service may be appropriate.

6. The examinees used in the sample must represent the population of interest. Because the value of the MH method at each score level is weighted by the number of examinees at that level, DIF displayed in very difficult items may go undetected. For these items, examinees in most of the score levels have a chance probability of a correct response regardless of group membership. DIF only occurs at the extreme end of the ability distribution. If examinees from that part of the distribution represent a small proportion of the sample, such DIF will go undetected. Oversampling of examinees at the higher end of the ability scale is needed to overcome the problem.

7. Combining score groups in the matching criterion may be useful for increasing the power of the MH statistic, but, in general, this practice should be avoided. When examinee ability distributions are similar for the majority and minority groups, this practice may lead to a modest increase in statistical power with little increase in error. When unequally distributed majority and minority groups are compared, a substantial increase in error may result, making the validity of the procedure questionable. Since the equal ability distribution condition is likely to exist for those groups from which it is easy to collect

larger samples (e.g., males and females), the overall utility of combining score groups must be considered questionable.

8. Practitioners should remain aware of the fact that items that have lower a-parameters (item discrimination parameters) are less likely to be identified by the MH method. If items with a-parameters below 0.6 are a small percentage of the items on a test, they can probably be discounted. If, by contrast, most of the items on the test are of this type, it may be appropriate to use a lower significance level when screening for these items. This condition would be typical of some certification tests.

9. The results cited from our studies make it clear that the MH method is not "blind" to non-uniform DIF. Nonetheless, it is insensitive to DIF in many items of this type. Appropriate screening for such items requires use of an additional method. This could include the variation on the MH statistic described in Mazor, Clauser, and Hambleton (in press).

At this time, our research program is continuing along several lines. Rogers and Swaminathan (1990) continue to investigate the utility of two-parameter logistic regression models. A main advantage of these models is that they can detect non-uniform DIF while having most of the advantages of the MH method. Kathy Mazor's dissertation research is focused on the utility of logistic regression models with multiple criterion measures for handling multidimensionality in the test data in the detection of DIF. Finally, we are beginning to apply DIF detection methods to the study of the equivalence of test translations. The problems of translating tests and establishing test score equivalence have become central concerns in the design of valid international comparative studies of educational achievement. Such studies

are becoming very important as national governments search for ways to improve the quality of education.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.

Angoff, W. H. (1982). The use of difficulty and discrimination indices in the identification of biased test items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 96-116). Baltimore, MD: John Hopkins University Press.

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

Clauser, B. E. (1993). Factors influencing the performance of the Mantel-Haenszel procedure in identifying differential item functioning. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991a). The influence of the criterion variable on the identification of differentially functioning items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 15(4), 353-359.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1991b). Examination of various influences on the Mantel-Haenszel statistic (Laboratory of Psychometric and Evaluative Research Report No. 210). Amherst, MA: University of Massachusetts, School of Education.

Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (in press). The effects of score group width on the Mantel-Haenszel procedure. Journal of Educational Measurement.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item difficulty of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach (Research Report No. 83-9). Princeton, NJ: Educational Testing Service.

Durovic, J. J. (1975). Definitions of test bias: A taxonomy and illustration of an alternative model. Unpublished doctoral dissertation, State University of New York, Albany.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed.; pp. 147-200). New York, NY: Macmillan Publishing Company.

Hambleton, R. K., Bollwark, J., & Rogers, H. J. (in press). Stability of Mantel-Haenszel DIF statistics across criterion measures and samples. Applied Measurement in Education.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of empirical and judgmental methods of detecting differential item functioning (Laboratory of Psychometric and Evaluative Research Report No. 231). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & Rogers, H. J. (1988). Design of an item bias review form: issues and questions (Technical Report). Albany, NY: New York Department of Education.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., & Rogers, H. J. (1991). Evaluation of the plot method for identifying potentially biased test items. In P. L. Dann, S. H. Irvine, & J. M. Collis (Eds.), Advances in computer based human assessment (pp. 307-330). Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

Holland, P. W., & Thayer, D. T. (1986, April). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

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

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 155-174). Vancouver, BC: Educational Research Institute of British Columbia.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 18, 209-225.

Kingston, N., Leary, L., & Wightman, L. (1988). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test (GMAC Occasional Papers). Princeton, NJ: Graduate Management Admission Council.

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

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-452.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (in press). Identification of non-uniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement.

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13(2), 127-143.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement, 14(2), 197-207.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2(1, 1-13.

Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Rogers, H. J., & Hambleton, R. K. (1985). IRT BIAS: A Fortran V program to compute several IRT DIF statistics. Amherst, MA: University of Massachusetts, School of Education.

Rogers, H. J., & Hambleton, R. K. (1989). Evaluation of computer simulated baseline statistics for use in item bias studies. Educational and Psychological Measurement, 49, 355-369.

Rogers, H. J., & Hambleton, R. K. (in press). MH: A Fortran V program to compute the Mantel-Haenszel statistic for detecting differential item functioning. Educational and Psychological Measurement.

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

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D. (1987). An experimental exploratory study of causes of bias in test items. Journal of Educational Measurement, 24(1), 97-118.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

Shepard, L. A. (1981). Identifying bias in test items. In B. F. Green (Ed.), New directions for testing and measurement: Issues in testing-coaching disclosure and ethnic bias. No. 11. San Francisco, CA: Jossey-Bass.

Shepard, L. A., 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.

Stricker, L. J. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. Applied Psychological Measurement, 6(3), 261-273.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21(1), 49-58.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore, MD: The Johns Hopkins University Press.

Veale, J. R. (1977). A note on the use of chi-square with "correct/incorrect" data to detect culturally biased items (Statistical Research: Education and the Behavioral Sciences, Technical Rep. No. 4). (Available from J. R. Veale, P.O. Box 4036, Berkeley, CA 94704)

Wright, B. D., Mead, R. J., & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model (Research Memorandum No. 22). Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.

Table 1

Agreement Between Methods in the Identification of DIF Test Items[1]

Test Item	IRT Area Method Sample 1	IRT Area Method Sample 2	Mantel-Haenszel Method Sample 1	Mantel-Haenszel Method Sample 2	Agreement
11	(0.354)[2]	0.645	17.49	20.56	
28	0.903	0.657	(0.38)	(0.00)	
30	0.736	0.701	(0.10)	(4.54)	
57	0.584	0.562	7.34	(4.94)	
60	(0.315)	(0.349)	8.08	7.30	
82	0.509	0.626	20.17	8.56	X
88	0.516	0.493	(2.90)	(0.46)	
92	0.686	0.916	(0.01)	(0.11)	
101	0.838	0.602	30.87	8.32	X
102	0.584	0.488	9.67	(0.36)	
107	0.567	0.581	11.43	11.2	X
110	0.465	0.694	13.03	17.37	X
122	0.945	0.789	21.11	14.00	X
128	0.617	0.732	19.50	16.27	X
129	0.477	0.941	(2.11)	(0.41)	
130	0.747	0.577	7.49	12.59	X

[1]Test items listed in the table were consistently identified as DIF by one or both methods.

[2]DIF statistics reported in brackets were not significant.

Table 2

Summary of Change of Internal Criterion Variable Study

Internal Criterion[1]	Number of Items Consistently Identified as DIF Regardless of Context	Number of Items Which Changed Classification When Context Changed
Math Items (n=27)	1	3
Reading Items (n=15)	2	4
Prior Knowledge Items (Factual) (n=37)	6	8
Prior Knowledge Items (No Clearly Best Answer) (n=12)	6	2
Charts Items (n=19)	1	7
Control Items 1 (n=30)	7	2
Control Items 2 (n=31)	6	6
Control Items 3 (n=30)	5	8

[1]Number of items in each internal criterion appears in brackets.

Table 3

Summary of Sample Size Results

Sample Size per Group	Largest p-difference Missed		Smallest p-difference Identified		Percentage of DIF Items Correctly Identified	
	Equal Dist.	Unequal Dist.	Equal Dist.	Unequal Dist.	Equal Dist.	Unequal Dist.
100	.23	.29	.14	.09	18%	9%
200	.17	.23	.07	.03	28%	24%
500	.08	.17	.07	.03	38%	31%
1000	.08	.15	.03	.01	61%	58%
2000	.04	.07	.02	.01	74%	64%

Table 4

Detection Rates for Non-Uniform DIF Items

Majority and Minority Ability Distributions	a-Difference[1] 1.0	.75	.50	.25
Equal				
Percentage of Items Identified With Full Sample	73%	73%	69%	60%
Percentage of Items Identified With Full Sample or Split Sample	93%	90%	84%	64%
Unequal				
Percentage of Items Identified With Full Sample	69%	66%	60%	48%
Percentage of Items Identified With Full Sample or Split Sample	89%	83%	76%	53%

[1]a is the item discrimination parameter in the three-parameter logistic model.

Table 5

Summary of Major MH DIF Research Study Findings

Study	Variables	Major Findings	Implications
Internal vs. External Criterion	Impact of choice of criterion-internal or external, on the results of DIF studies	Even when correlations between criterion measures are only moderate, the results of DIF studies are very similar.	Use of the more convenient internal criterion (i.e., total test score or some subtest) can be defended.
Influence of Criterion Variable	Impact of choice of internal criterion-variable - compared results of matching on total test score with matching on subtest score.	Choice of internal criterion has a substantial impact on item classifi - cations - many items change classification when the criterion changes.	Practitioners should use caution when interpreting MH results. Item classifications as DIF or not-DIF are not absolute but depend on context.
Sample Size	Sample size - varied sample size from 100 to 2000 examinees per group. Compared detection rates across sample sizes.	As sample size increases, identification rate increases. False positive rates are low.	With 2000 examinees per group, MH is sensitive to very small b differences. 1000 or 500 per group appears sufficient. 200 per group will result in more items missed, but in some circumstances may be sensitive enough. 100 per group cannot be recommended.

Table 5 continued:

Study	Variables	Major Findings	Implications
Various Influences	Examined the characteristics of DIF items which were most likely to be identified with MH statistic.	Greater b-differences associated with higher identification rates. Larger a-values associated with higher identification rates. Very difficult items (high b-values) more likely to be missed.	Poorly discriminating items may be difficult to detect. Very difficult items may also be missed.
Score-Group Width	Examined whether fewer than the maximum number of score groups improves identification rates.	Reducing the number of score groups (and thereby increasing score group width) does not result in significant changes in detection rates for equal ability distributions. However, for unequal distributions, identification rates improve somewhat, but false positive rates increase substantially as well.	Decreasing the number of score groups should be done with caution, especially with groups of differing ability distributions, as inflated type I error may be expected.

Table 5 continued:

Study	Variables	Major Findings	Implications
Non-Uniform DIF	Assessed whether the MH method and a variation of the standard method would identify non-uniform DIF items.	The MH method is able to identify many non-uniform DIF items. The variation (splitting the sample into high and low performing samples and re-running the method on each group) increases identification rate further.	The MH method is sensitive to some types of non-uniform DIF. If non-uniform DIF is a concern, detection rates can be improved by a variation of the standard method.

Figure 1. Plot of 1P average absolute SRs against point-biserial correlations.

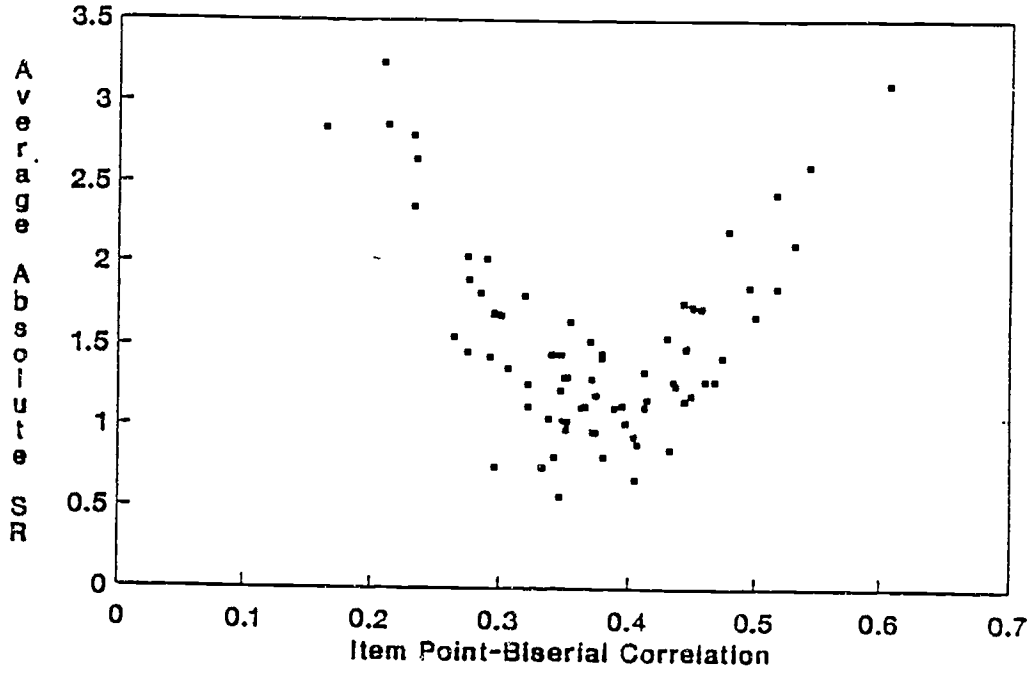

Figure 2. Plot of 2P average absolute SRs against point-biserial correlations.

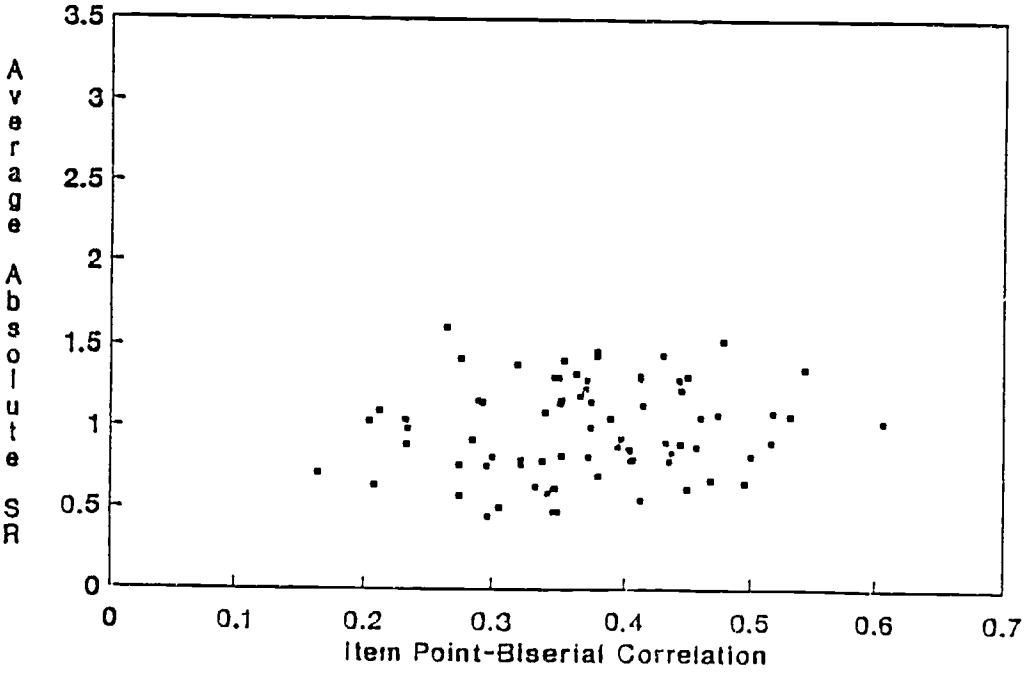

Figure 3. Graphical illustrations showing the difference between uniform and non-uniform DIF.