DOCUMENT RESUME

ED 372 097 TM 021 769

AUTHOR Pommerich, Mary; And Others
TITLE The Performance of the Mantel-Haenszel DIF Statistic When Comparison Group Distributions Are Incongruent.
PUB DATE Apr 94
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Comparative Analysis; *Item Bias; *Performance; Ratios (Mathematics); Research Methodology; Sample Size; *Scores; Simulation; *Statistical Distributions; Statistical Studies
IDENTIFIERS Contingency Tables; *Mantel Haenszel Procedure

ABSTRACT

The functioning of two population-based Mantel-Haenszel (MH) common-odds ratios was compared. One ratio is conditioned on the observed test score, while the other is conditioned on a latent trait or true ability score. When the comparison group distributions are incongruent or nonoverlapping to some degree, the observed score represents different levels of latent ability across the comparison groups, raising a question as to the effectiveness of observed score matching under conditions that could influence performance of the MH statistic in the identification of differential item functioning (DIF). The current study varies from typical simulation methodology in that the sample sizes are assumed to be infinite, and the observed score MH common-odds ratio is computed from the expected cell frequencies of the 2x2 contingency tables. A MH common-odds ratio based on latent ability is computed to define a measure of true DIF. Under all conditions examined, the observed score MH performed similarly to the latent ability MH. This provides reassurance in conditioning on the observed score when the MH statistic is applied to large finite samples with comparison groups that are not completely overlapping. Two figures and five tables are included. (Contains 6 references.) (SLD)

The Performance of the Mantel-Haenszel DIF Statistic When Comparison Group Distributions are Incongruent[1]

Mary Pommerich
University of North Carolina at Chapel Hill
Judith A. Spray and Cynthia G. Parshall
American College Testing

Abstract

This study was conducted to compare the functioning of two population-based Mantel-Haenszel (MH) common-odds ratios. One ratio conditioned on the observed test score, while the other conditioned on a latent trait or true ability score. Commonly, the observed test score is used in the calculation of the MH statistic as a surrogate for the true but unknown latent ability of each examinee. When the comparison group distributions are incongruent, or non-overlapping to some degree, observed score represents different levels of latent ability across the comparison groups; a question remains as to the effectiveness of observed score matching under conditions that could influence the performance of the MH statistic in the identification of differential item functioning (DIF).

In similar studies, simulation methodology has been employed to perform replications of DIF calculations using finite samples drawn from two comparison groups or populations. Typically, the sample sizes of the groups are manipulated, and the effect of sample size is observed on the detection of simulated DIF. The current study varies from the typical simulation methodology in several important ways. First, the sample sizes from both comparison groups were assumed to be infinite, and the observed score MH common-odds ratio was computed from the expected cell frequencies of the 2 x 2 contingency tables. Second, a MH common-odds ratio based on latent ability was computed to define a measure of true DIF. The latent ability MH provides a standard of comparison for the observed score MH. The use of these population-based MH common-odds ratios allowed an evaluation of

[1]Paper presented at the Annual Meeting of the American Educational Research Association, April 4-8, 1994, New Orleans, LA.

the MH statistic as sample sizes approached infinity, eliminating the sample size effect from the study.

The performance of the population-based MH common-odds ratios on tests of moderate and high difficulty was evaluated for combinations of percentage of distributional overlap, test length, occurrence of DIF in test items, and relative proportion of examinees in the comparison groups. Under all of the conditions examined, the observed score MH common-odds ratio performed similarly to the latent ability MH, even with moderately congruent distributions. Manipulations of test length, the occurrence of DIF, and the proportional mix of the comparison distributions did not produce substantial differences between the two population-based common-odds ratios on tests of moderate and high difficulty, under fairly incongruent distributions. This provides reassurance in conditioning on observed score when the MH statistic is applied to large finite samples with comparison group distributions that are not completely overlapping.

The Performance of the Mantel-Haenszel DIF Statistic
When Comparison Group Distributions are Incongruent

Mary Pommerich
University of North Carolina at Chapel Hill
Judith A. Spray and Cynthia G. Parshall
American College Testing

A common approach to the detection of differential item functioning (DIF) in two comparison groups is to employ the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959) to flag test items where DIF might be problematic. Under this approach, the performance of a focal group on an item of interest (the studied item) is compared to the performance of a reference group, where the reference group provides a standard for comparison. The two groups are typically matched on some criterion—often total test score—so that if DIF occurs, a distinction can be made between a simple difference in the relative ability of unmatched comparison groups (a measure of impact) versus true differential functioning attributable to the item. Holland and Thayer (1988) assert that use of a matching criterion ensures that only comparable members of the comparison groups are employed, where comparability implies identity of examinees on measured characteristics that are strongly related to performance on the studied item.

The Mantel-Haenszel Statistic

Once the groups are matched on some criterion variable, the comparable examinees can be placed into s 2×2 tables of group-by-item response, where s equals the number of levels of the matching variable. If s indexes each observed score category of a k-item test, with $s = 0, 1, ..., k$, then one 2×2 table for a given item within score category s can be represented as

	Correct	Incorrect	Total
Reference	R_R	W_R	N_R
Focal	R_F	W_F	N_F
Total	R_s	W_s	N_s

where R_R, R_F, and R_s are the frequencies of correct responses to the item in the reference group, the focal group, and the combined group, respectively, at s; W_R, W_F, and W_s are the frequencies of incorrect responses to the item in the reference, focal, and combined groups, respectively, at s; and N_R, N_F, and N_s are the total number of examinees within the reference, focal, and combined groups, respectively, at s. The tabled information is employed in the computation of a common-odds ratio estimator, given by

$$MH = \frac{\sum_{s=0}^{k} \frac{R_R W_F}{N_s}}{\sum_{s=0}^{k} \frac{R_F W_R}{N_s}} . \tag{1}$$

The MH index can also be given in terms of the proportion of correct responses within each group:

$$MH = \frac{\sum_{s=0}^{k} P_R Q_F \frac{G_R \cdot G_F}{G_s}}{\sum_{s=0}^{k} P_F Q_R \frac{G_R \cdot G_F}{G_s}} , \tag{2}$$

where P_R and P_F are the proportions correct for the reference and focal groups at s, respectively; Q_R and Q_F are defined as $(1 - P_R)$ and $(1 - P_F)$, respectively; G_R and G_F are the relative frequencies of the reference and focal groups at s; and G_s is the total relative frequency of the reference and focal groups at s. Specifically,

$$P_R = \frac{R_R}{N_R} \qquad P_F = \frac{R_F}{N_F}, \tag{3}$$

$$G_R = \frac{N_R}{\sum_{s=0}^{k} N_R} \qquad G_F = \frac{N_F}{\sum_{s=0}^{k} N_F},$$

and

$$G_s = \frac{N_s}{\sum_{s=0}^{k} N_s} .$$

The value of the MH statistic indicates, on the average, the extent to which it is more (or less) likely that a member of the reference group answered the item correctly than did a comparable member of the focal group. If there is no differential functioning between the comparison groups on that item, the value of the MH statistic is *1.0*. For an item with DIF, the MH value will be greater than *1.0* when the item favors the reference group and less than *1.0* when the item favors the focal group. A formal hypothesis for the common-odds ratio of an item is represented by the null hypothesis

$$H_0: \frac{R_R}{W_R} = \frac{R_F}{W_F}. \tag{4}$$

When $MH = 1$, the null hypothesis is met; when $MH \neq 1$, the alternative hypothesis holds:

$$H_1: \frac{R_R}{W_R} = MH\frac{R_F}{W_F}. \tag{5}$$

When the observed score is used as the matching criterion, it is questionable whether the MH statistic functions well when the distributions of the comparison groups are incongruent or non-overlapping to some degree. As observed by Spray and Miller (1992)[2], conditioning on the observed test score appears to be appropriate provided the observed test score accurately reflects a comparable level of the measured trait for the populations of interest. Problems may arise when identical values of the observed test score represent different levels of ability across groups, such as when the conditional distributions of ability given observed score are different, or incongruent, for the focal and reference groups. If the MH is unstable under incongruent distributions and performs poorly, then its application may be inappropriate under such conditions. This study was conducted to evaluate the effectiveness of observed score matching when comparison group distributions are incongruent, under a variety of analysis conditions.

[2]This study differs from a previous study (Spray & Miller, 1992) that looked at similar effects of incongruent ability distributions on the MH statistic. The present study employs analytical methods and does not rely on computer simulation results with finite samples. Also, the computation of the observed score MH value (computed from expected cell frequencies) in the current study utilizes a correct algorithm. Although the Spray and Miller paper presented the results of the simulations accurately, a section that attempted to show what would happen as cell sample sizes approached infinity was based on an incorrect computing formula.

The performance of the MH statistic under incongruent ability distributions was studied from a theoretical perspective by Zwick (1990). When the matching variable was total test score (excluding the studied variable), Zwick concluded that the MH null hypothesis (Equation 4) would not be satisfied if the ability distributions were not identical for both groups, even where all of the items were free of DIF. Further, where the comparison distributions were incongruent, the MH would show DIF favoring the group with higher ability. When the studied item was included in the matching criterion, Zwick determined that in general the MH null hypothesis would not hold when there was no DIF, and that it was possible for the MH to show DIF favoring either of the comparison groups when ability distributions were incongruent. Specifically, the MH would show DIF favoring the higher ability group when the probability of getting an item correct (given ability, score, and group membership) was monotonically increasing with ability. The MH would show DIF favoring the lower ability group when the probability of getting an item correct was monotonically decreasing with ability.

Zwick's (1990) general conclusion was confirmed by Schulz, Perlman, Rice, and Wright (in press) in their study of MH procedures for assessing DIF, but in some instances where directional favoring did occur under incongruent distributions, the MH favored the ability group in the opposite direction as that suggested by Zwick.

Method

DIF Indices

The MH statistic given in Equations 1 and 2 is defined in terms of observed test score, leading to potential inaccuracies in the resulting value when the observed test score is not a reflection of the underlying latent ability of the test taker. When matching examinees across comparison groups, conditioning on latent ability of the examinee—or true test score—is preferable to conditioning on observed score. A MH value based on latent ability yields a population definition of the common-odds ratio, and represents a true but unknown measure of DIF in an item.

For this study, two population-based MH common-odds ratios were defined. First, the sample sizes from both comparison groups were assumed to be infinite, and a MH common-odds ratio conditioned on observed score was computed from the expected cell frequencies of the contingency tables for the score categories. Second, a MH common-odds ratio based on

latent ability matching was computed to provide a standard of comparison for the observed score MH. Computation of these two MH common-odds ratios ensured that simulation of item response data was unnecessary to the study, as they do not require samples for their calculations. Accordingly, the question of appropriate sample size to include in the computations was not an issue in this study.

Observed Score MH

A population-based MH common-odds ratio conditioned on observed score can be formed by using the expected cell frequencies in Equation 1, or by the expected cell proportions in Equation 2. For this study, the observed score common-odds ratio is defined as

$$MH_X = \frac{\sum P_R(U=1|X)[1 - P_F(U=1|X)]\frac{F_F(X)F_R(X)}{F^*(X)}}{\sum P_F(U=1|X)[1 - P_R(U=1|X)]\frac{F_F(X)F_R(X)}{F^*(X)}}, \tag{6}$$

where $P_R(U=1|X)$ and $P_F(U=1|X)$ are the probabilities of a correct response given X, in the reference and focal groups, respectively; and $F_R(X)$, $F_F(X)$, and $F^*(X)$ are the expected observed score frequencies of the reference, focal, and combined groups, respectively. The probability of a correct response in the reference group, given observed score, is computed by

$$P_R(U=1|X) = \frac{\int_{-\infty}^{\infty} P(U=1|\theta)P(Y|\theta)g(\theta)d\theta}{\int_{-\infty}^{\infty} P(X|\theta)g(\theta)d\theta}, \tag{7}$$

where

U = item score for the studied item,

Y = sum of the item scores excluding the studied item,

and

X = Y + U .

A similar definition holds for the focal group. The expected observed score frequencies are calculated from

$$\int_{\theta} h(X|\theta) g(\theta)\, d\theta\,, \tag{8}$$

where $h(X|\theta)$ is the compound binomial probability of observing X, given θ. It is calculated using a recursive technique given by Lord and Wingersky (1984).

Latent Ability MH

The common-odds ratio conditioned on latent ability, MH_θ, is defined as

$$MH_\theta = \frac{\int_{-\infty}^{\infty} P_R(\theta)\, Q_F(\theta)\, \frac{g_F(\theta) g_R(\theta)}{g^*(\theta)}\, d\theta}{\int_{-\infty}^{\infty} P_F(\theta)\, Q_R(\theta)\, \frac{g_F(\theta) g_R(\theta)}{g^*(\theta)}\, d\theta}\,. \tag{9}$$

Note that the proportions correct and incorrect (P_R, P_F, Q_R, Q_F) at each score category from the sample estimator of the common-odds ratio given in Equation 2 are replaced with probability functions of θ, the latent ability variable. The probabilities of correct response, $P_R(\theta)$ and $P_F(\theta)$, are given by the unidimensional three-parameter logistic item response function,

$$P(\theta) = c + \frac{(1-c)}{1 + e^{-1.7a(\theta - b)}}\,, \tag{10}$$

while $Q_R(\theta) = 1 - P_R(\theta)$ and $Q_F(\theta) = 1 - P_F(\theta)$. Latent ability, θ, is assumed to be a continuous random variable with known density functions, defined as

$$g_R(\theta) = \frac{1}{\sqrt{2\pi\sigma_R^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \mu_R)^2}{\sigma_R^2}\right) \tag{11}$$

and

$$g_F(\theta) = \frac{1}{\sqrt{2\pi\sigma_F^2}} \exp\left(-\frac{1}{2}\frac{(\theta-\mu_F)^2}{\sigma_F^2}\right) \tag{12}$$

in the reference and focal groups, respectively. The combined group density is computed using

$$g^*(\theta) = \alpha g_F(\theta) + (1-\alpha) g_R(\theta), \tag{13}$$

where α represents the relative proportion of examinees contained in the focal group, with $0 \leq \alpha \leq 1$.

Analysis Conditions

Degree of Distributional Incongruence

Of interest in the study was the performance of the population-based MH common-odds ratios when the abilities of the comparison groups were discrepant, or incongruent to differing degrees, under various conditions. The primary question was whether matching on latent ability or observed score would yield consistent MH values when the overlap between the comparison distributions was not complete. A measure of the degree to which the two distributions were incongruent was given by the percentage of overlap of the areas under the density functions of the comparison groups. This measure allowed for an infinite number of combinations of distributions to be mapped to a simple scalar between 0.0 (signifying no overlap, or total incongruence) and 1.0 (signifying complete overlap, or total congruence). The measure was defined as

$$\text{PERCENT } OVERLAP = \int_{-\infty}^{\infty} MIN\,[g_R(\theta), g_F(\theta)]\,d\theta. \tag{14}$$

Throughout the study, the degree of overlap was varied by manipulating the focal group distribution. In the computation of the MH common-odds ratio, the reference group was always drawn from a normal distribution with mean *0* and variance *1*, while the focal group was drawn from the varying distributions *N(0,1)*, *N(0,.5)*, *N(-1.5,1)*, *N(-1.5,.5)*, *N(-3,1)*, and *N(-3,.5)*. The corresponding degrees of overlap (listed in Table 1) ranged from complete

congruence under a focal group distribution of $N(0,1)$ to virtually complete incongruence under a focal group distribution of $N(-3,.5)$.

Insert Table 1 About Here

Parameter Generation

The IRT parameters for the focal and reference groups were generated so that the a parameters were uniformly distributed between $.5$ and $.75$ and the c parameters were uniformly distributed between $.05$ and $.10$. Two ranges were examined for the b parameters: in Experiment 1 the b parameters were constricted within the range of $-.5$ to $.5$, while in Experiment 2, the b parameters ranged from 1.0 to 2.0. This yielded a homogeneous test of moderate difficulty for both groups in Experiment 1 and a high difficulty test for the comparison groups in Experiment 2, particularly for the focal group.

Under the condition of no DIF, the generated parameters were set equal in the focal and reference groups across all items. Thus, while the parameter values varied within the specified ranges across items, there was no parameter variation across the comparison groups. Under the condition of DIF, a small amount of DIF favoring the reference group was induced in the b parameter of one item by setting $b_F = b_R + .3$ for that item. As in the no DIF condition, the a and c parameters remained equal across the two groups for the studied item. For the items in which no DIF was induced, all parameters were set equal for each item across groups, while varying across items.

Test Length and Ratio of Examinees

Two additional conditions were manipulated throughout the two experiments—the test length and the ratio of focal to reference group examinees used in creating the combined group density. The test length was set at 20, 40, or 80 items. The ratio was set at $1:10$ or $1:1$, so that $\alpha = 1/11$ or $\alpha = 1/2$.

The final experimental design was a $6 \times 3 \times 2 \times 2$ factor experiment with six levels of overlap, three levels of test length, two levels of DIF (DIF or no DIF), and two levels of the ratio of focal to reference group examinees. This produced a total of 72 research conditions within each of the two experiments.

Results

The observed score MH common-odds ratio and the latent ability MH common-odds ratio were computed for all combinations of the experimental conditions. The MH common-odds ratio conditioned on latent ability provides a standard of comparison for the performance of the MH common-odds ratio conditioned on observed score. Of interest in the study was the performance of the observed score MH common-odds ratio under the manipulated conditions, relative to the corresponding latent ability MH common-odds ratio.

Because the MH common-odds ratios used in this study were by definition sample-free in their computation, the resulting data consisted of effects that were considered to be actual parameter values rather than estimates. Inferential analyses of these MH values were not deemed appropriate, given the population status of the defined common-odds ratios. Hence, only descriptive statistics for the common-odds ratios are reported in this paper.

Experiment 1

The descriptive statistics for the experiment in which the *b* parameters were restricted to the moderately difficult range (-.5 to .5) are presented in Tables 2 and 3. Table 2 gives the results for a ratio of *1:1*, while Table 3 gives the results for a ratio of *1:10*. Within each table, information is given on the observed score MH common-odds ratio (MH_X) averaged across items and the standard deviation of MH_X. (The values are reported in the columns headed Ave MH_X and SD MH_X.) Under the condition of no DIF (DIF=N), all items were included in the computation of these statistics; under the condition of DIF (DIF=Y), the item containing DIF was excluded from the computation of the average and standard deviation of MH_X. For the DIF induced items alone, MH_X and the latent ability MH common-odds ratio (MH_θ) are reported for that item (given in the columns labeled MH_X and MH_θ). The latent ability MH is only reported for the DIF condition because under the condition of no DIF, the value was always *1.0* for all items. The difference between MH_θ and MH_X was also computed (reported in the column labeled θ-X). Also given in the tables are the reliability of each test for both the reference and focal groups (listed in the columns labeled r_R and r_F, respectively) and the difficulty of the DIF-induced item for the reference group (reported in the column headed b_R).

Examination of the two tables show parallel results for the MH common-odds ratios across the two ratios of relative group size; thus only results from Table 2 are discussed. The

similarity of results implies that the ratio of examinees is not a critical factor in determining the value of the MH common-odds ratios; the relative size of the comparison groups appears irrelevant to the outcome.

Insert Tables 2 and 3 About Here

No DIF Condition

Under the condition of no DIF, the observed score MH averaged across all items (Ave MH_X) consistently yielded values around *1.0*, as predicted, for all degrees of overlap and all test lengths. The standard deviation of MH_X (SD MH_X), however, showed an increase in variability in the MH_X across items as the distributions became more incongruent, particularly with *20* item tests. As the test length increased within each category of distributional incongruence, the variability across items decreased. The trend in variability demonstrated across categories of distributional incongruence here indicates that although the average MH_X was *1.0*, more items are likely to be falsely identified as displaying DIF as the degree of distributional incongruence increases. While greater numbers of items would be less likely to result in false positives, the test lengths employed in the study do not appear to be critical to the functioning of the observed score MH common-odds ratio.

DIF Condition

When DIF was induced in one item, the average MH_X (excluding that DIF item) again fell consistently around *1.0*, although slightly below the predicted value of *1.0*. The occurrence of DIF in one item appeared to affect the remaining items by pulling their expected value below *1.0*. The degree of variability in the average MH_X followed a pattern similar to that found under the no DIF condition across differing test lengths.

For the single DIF item, both MH_X and MH_θ consistently showed DIF favoring the reference group, with a larger value for MH_θ. The absolute value of the difference between MH_θ and MH_X (θ-X) as a function of percent overlap is plotted in Figure 1. The difference between the latent ability and observed score MH values within each test length remained fairly constant with increasing distributional incongruence, up to the point where the group means were three standard deviations apart (percent overlap < *.15*). Across the three test lengths, the θ-X difference also remained close, up to the point where the overlap between group means was less than *.15*.

While MH_θ remained fairly constant across the conditions of incongruence, the observed degree to which the item favored the reference group decreased, with MH_X approaching *1.0*, as the distributions became more incongruent. This trend was unexpected in light of Zwick's (1990) prediction that the MH would produce a conclusion of DIF favoring the reference group when the distributions were ordered with a higher mean for the reference distribution. The logical assumption would be that the degree of favoring for the higher mean group would increase rather than decrease as distributions become more incongruent. However, the observed similarities between the MH_X and MH_θ values suggest that distributional incongruence is not likely to lead to inaccurate assessments of the direction and magnitude of DIF under the given conditions, up to a minimal degree of overlap between the comparison distributions.

Test Reliability and Item Difficulty

In addition to the MH common-odds ratios, the reliability of each test was computed for both the reference group (r_R) and focal group (r_F). Reliabilities for the reference group remained high throughout the full range of overlap, while reliabilities for the focal group fell as low as *.17* under the *20*-item DIF condition within the most incongruent of the comparison distributions. Despite the very poor reliability that often occurred within the focal group, the MH common-odds ratios did not appear to be adversely affected. When there was no DIF, the observed score MH common-odds ratio averaged across all items (Ave MH_X) was very close to *1.0*, even in situations where focal group reliability was unacceptably low. Variability of the average MH_X (SD MH_X) did increase inversely with reliability, indicating that in the case of a low reliability test, a false positive identification of DIF would be more likely to occur than with a highly reliable test. When DIF was induced, the fluctuations in MH_X were not consistent with the variations in reliability. The reliability of the test alone does not appear to be very influential in determining the degree of DIF observed in items. Under conditions of moderate overlap, the observed score MH performs similarly to the latent ability MH regardless of the reliability of the test.

One final consideration was the effect of the difficulty of the item on the observed score MH common-odds ratio. For this experiment, the item difficulty parameters were sampled from a constricted range yielding a homogeneous test of medium difficulty. In the tables, the difficulty parameters of the DIF items for the reference group are reported in the

column headed b_R. It appears that the MH_X value may have been confounded somewhat by the degree of difficulty in the DIF-induced item. As distributional incongruence increased, high negative values of difficulty tended to have the higher values of MH_X, while the high positive values of difficulty had the lower values of MH_X. The degree of DIF may be controlled somewhat by the difficulty of the item of interest. This trend is difficult to characterize because the range of values for item difficulty was restricted between *-.5* and *.5*. It is possible that more discrepant values of MH_X would occur where item difficulty is allowed a wider range of values.

Experiment 2

The second experiment differed from the conditions of Experiment 1 in that the item difficulties ranged from *1.0* to *2.0*. The range was restricted in Experiment 2 to create a difficult homogeneous test, one that was particularly difficult for the focal group. The descriptive statistics for this experiment are presented in Tables 4 and 5. Table 4 gives the results for a ratio of *1:1*, while Table 5 gives the results for a ratio of *1:10*. Examination of the two tables shows very similar results across the two ratio conditions, therefore only the results from Table 4 will be discussed. The information reported in Table 4 is identical to that discussed with Table 2 in Experiment 1.

Insert Tables 4 and 5 About Here

No DIF Condition

Under the condition of no DIF, the average observed score MH (Ave MH_X) values were very close to the hypothesized value of *1.0*. The variability of the observed score common-odds ratio increased as the distributions became more incongruent, with an obvious jump in the amount of variability demonstrated at a distance of *3.0* standard deviations between distribution means. Variability also increased as the test length decreased. The same trend in variability across test length was observed in Experiment 1 (see Table 2), but the degree of variability in Experiment 2 was consistently greater than that of Experiment 1. The more difficult test yielded less consistent values of MH_X than the less difficult test when no DIF occurred in the test items.

DIF Condition

When DIF was induced in one item, Ave MH_X (excluding the DIF item) also fell close to *1.0*, with the degree of variability showing a pattern similar to that of the no DIF situation. The inducement of DIF in one item did not affect the value of the observed score common-odds ratio in the non-DIF items. Both MH common-odds ratios (MH_X and MH_θ) showed DIF favoring the reference group in all cases with the exception of an MH_X falling below *1.0* under a *20*-item test within the most incongruent condition. The degree to which MH_X favored the reference group appeared to decrease, however, as the comparison distributions displayed less overlap. A similar tendency was noted in Experiment 1, where item difficulty was constrained within a moderate range.

The absolute value of the difference between MH_θ and MH_X (θ-X) as a function of percent overlap is plotted in Figure 2. The difference between latent ability and observed score MH values within *80* item tests remained fairly constant with the increasing distributional incongruence. For test lengths of *20* and *40* items, the difference in the MH common-odds ratios varied across the increasing distributional incongruence. Across the three test lengths the θ-X difference remained fairly close, beginning to diverge where percent overlap was less than *.37*. The difference between the two common-odds ratios appeared to grow larger as the distributions became more incongruent, although the trend was not consistent. While MH_θ remained fairly constant across the conditions of incongruence, the observed degree to which the item favored the reference group decreased, with MH_X approaching or falling below *1.0* as the distributions became more incongruent. Only under conditions of very extreme incongruence with test lengths of *20*-items does it appear that the observed score MH common-odds ratio would give a value showing favor in a direction that did not correspond to the latent ability MH value.

Across the two experimental conditions, the observed score MH common-odds ratio (MH_X) in Experiment 2 was consistently less than MH_X in Experiment 1, until the distributions were three standard deviations apart. The discrepancy between the latent ability and observed score MH values (θ-X) was generally greater within the very difficult test than within the moderately difficult test. This demonstrates that under a very difficult test, false identification of DIF is probably more likely to occur than under a moderately difficult test.

Test Reliability and Item Difficulty

When the reliabilities of the test were examined for each group, the reliability for the reference group remained consistently high as the distributions became more incongruent, while the reliability for the focal group grew very poor as the degree of overlap lessened. Focal group reliability reached a minimum of *.02* with a *20*-item test under the most incongruent condition. Focal group reliabilities were as low as *.20* when the distributions were *1.5* standard deviations apart, yet the functioning of the observed score MH common-odds ratio did not appear to be affected by the reliability at this degree of incongruence. As concluded in Experiment 1, reliability does not seem to be influential in the functioning of the observed MH common-odds ratio. Likewise, while a longer test is generally preferable, the actual test length showed only a minor effect on the observed score MH value.

Finally, examination of the item difficulty parameters for the DIF items showed the possibility of item difficulty confounding the resulting observed score MH value. As witnessed in the moderately difficult test situation, items with lower values of item difficulty tended to have larger values of MH_X, while more difficult items tended to have lower values of MH_X. The magnitude of the observed score MH common-odds ratio in an item may be affected by the difficulty of that item, leading to the potential misclassification of DIF. The relationship between item difficulty and magnitude of the observed score MH was not consistent across varying values of item difficulty, however, which indicates that item difficulty might work in combination with the other conditions to determine the resulting MH value.

Conclusion

Of primary interest in this study was the performance of the observed score MH common-odds ratio when the comparison distributions were incongruent. The results provide reassurance for using an observed score MH common-odds ratio with large finite sample sizes despite lack of complete overlap in the focal and reference group distributions. In both Experiment 1 and Experiment 2, the population-based observed score MH performed similarly to the latent ability MH in both DIF and non-DIF situations even to the point where distributions were as far as *1.5* standard deviations apart. Only when the degree of congruence fell below *.37* (with group mean differences of *3.0* standard deviations) did the

population-based observed score MH become distorted, particularly when all test items were very difficult.

Under all of the conditions examined, the population-based observed score MH common-odds ratio demonstrated great stability even with moderately congruent distributions. Test length and test reliability within groups did not play a critical role in determining the value of the MH. While greater numbers of items provided less variable results, the prevailing impression was that the test lengths examined were largely irrelevant to the outcome. Similarly, even with reliabilities as low as *.20*, the observed score MH performed well, excluding the conditions with the difference of *3.0* standard deviations.

If the stability of an observed score MH statistic under incongruent distributions in large finite samples is of concern, the results of this study indicate that matching on observed score to compute the value is a legitimate practice. The correspondence between the observed score MH common-odds ratio (MH_X) and the latent ability MH common-odds ratio (MH_θ) provides this assurance, as the value matched on latent ability is an indicator of true DIF. Even under conditions of fairly discrepant distributions the MH utilizing matching on observed score yields stable and consistent results.

References

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel technique. In H. Wainer & H.I. Braun (Eds.), *Test validity* (Chapter 9, pp. 129-146). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 453-461.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Schultz, E.M., Perlman, C., Rice, W.K., & Wright, B.D. (in press). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In G. Englehard and M. Wilson (Eds.) *Objective measurement: Theory into practice.* Volume 3. Norwood, NJ: Albex.

Spray, J.A., & Miller, T.R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent* (ACT Research Report Series No. 92-1). Iowa City, IA: American College Testing.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, *15*, 185-197.

Table 1

Percentage of overlap of the focal and reference distributions; the reference group is always distributed N(0,1).

Focal Mean	Focal Variance	Percent Overlap
0.0	1.0	1.0000
0.0	0.5	0.8339
-1.5	1.0	0.4532
-1.5	0.5	0.3707
-3.0	1.0	0.1336
-3.0	0.5	0.0774

Table 2

Experimental results for moderate difficulty *b* parameters (-.5 to .5) and *1:1* ratio of examinees, with observed score MH (MH_X), latent ability MH (MH_θ), MH_θ-MH_X (θ-X), test reliabilities for reference group (r_R) and focal group (r_F), and item difficulty for reference group (b_R).

%Overlap	#items	DIF	Ave MH_X[a]	Sd MH_X[a]	MH_X[b]	MH_θ[b]	θ-X	r_R	r_F	b_R
1.0000	20	N	1.000	0.000	-	-	-	0.777	0.777	-
		Y	0.985	0.001	1.311	1.333	0.022	0.800	0.800	-0.33
	40	N	1.000	0.000	-	-	-	0.883	0.883	-
		Y	0.993	0.001	1.275	1.320	0.045	0.880	0.880	0.09
	80	N	1.000	0.000	-	-	-	0.938	0.938	-
		Y	0.996	0.000	1.386	1.463	0.077	0.940	0.940	0.37
0.8339	20	N	1.000	0.003	-	-	-	0.805	0.695	-
		Y	0.983	0.003	1.391	1.451	0.060	0.784	0.665	0.10
	40	N	1.000	0.002	-	-	-	0.882	0.803	-
		Y	0.992	0.002	1.381	1.435	0.054	0.882	0.804	0.06
	80	N	1.000	0.001	-	-	-	0.941	0.897	-
		Y	0.996	0.001	1.367	1.430	0.063	0.939	0.894	0.37
0.4532	20	N	1.001	0.042	-	-	-	0.784	0.700	-
		Y	0.990	0.034	1.252	1.400	0.148	0.787	0.692	0.24
	40	N	1.000	0.021	-	-	-	0.883	0.826	-
		Y	0.993	0.020	1.300	1.347	0.047	0.884	0.826	-0.48
	80	N	1.000	0.012	-	-	-	0.939	0.910	-
		Y	0.998	0.011	1.216	1.297	0.081	0.937	0.905	0.15
0.3707	20	N	1.002	0.051	-	-	-	0.794	0.533	-
		Y	0.979	0.043	1.513	1.463	-0.050	0.787	0.548	-0.47
	40	N	1.000	0.023	-	-	-	0.885	0.706	-
		Y	0.991	0.026	1.434	1.464	0.030	0.880	0.682	-0.25
	80	N	1.000	0.014	-	-	-	0.937	0.818	-
		Y	0.997	0.014	1.244	1.302	0.058	0.941	0.832	-0.33
0.1336	20	N	1.002	0.102	-	-	-	0.782	0.425	-
		Y	1.001	0.075	1.000	1.451	0.451	0.781	0.431	0.39
	40	N	0.999	0.059	-	-	-	0.882	0.573	-
		Y	0.992	0.064	1.447	1.462	0.015	0.885	0.559	-0.19
	80	N	0.999	0.039	-	-	-	0.939	0.727	-
		Y	0.998	0.037	1.136	1.370	0.234	0.939	0.725	0.41
0.0774	20	N	1.006	0.129	-	-	-	0.791	0.191	-
		Y	0.998	0.138	1.126	1.454	0.328	0.786	0.165	0.24
	40	N	1.001	0.104	-	-	-	0.885	0.315	-
		Y	0.990	0.085	1.256	1.340	0.084	0.885	0.336	-0.48
	80	N	0.998	0.059	-	-	-	0.937	0.499	-
		Y	0.997	0.057	1.023	1.319	0.296	0.938	0.475	0.39

a Computed from all items when DIF=N, excludes the DIF item when DIF=Y
b Computed on DIF item only

Table 3

Experimental results for moderate difficulty *b* parameters (-.5 to .5) and *1:10* ratio of examinees, with observed score MH (MH_X), latent ability MH (MH_θ), MH_θ-MH_X (θ-X), test reliabilities for reference group (r_R) and focal group (r_F), and item difficulty for reference group (b_R).

%Overlap	#items	DIF	Ave MH_X[a]	Sd MH_X[a]	MH_X[b]	MH_θ[b]	θ-X	r_R	r_F	b_R
1.0000	20	N	1.000	0.000	-	-	-	0.788	0.788	-
		Y	0.986	0.001	1.287	1.319	0.032	0.791	0.791	-0.08
	40	N	1.000	0.000	-	-	-	0.885	0.885	-
		Y	0.993	0.001	1.302	1.340	0.038	0.881	0.881	0.22
	80	N	1.000	0.000	-	-	-	0.938	0.938	-
		Y	0.996	0.000	1.332	1.385	0.053	0.939	0.939	0.38
0.8339	20	N	1.000	0.003	-	-	-	0.784	0.665	-
		Y	0.985	0.004	1.342	1.385	0.043	0.791	0.674	-0.35
	40	N	1.000	0.002	-	-	-	0.882	0.804	-
		Y	0.993	0.002	1.311	1.346	0.035	0.883	0.806	0.07
	80	N	1.000	0.001	-	-	-	0.938	0.893	-
		Y	0.996	0.001	1.360	1.401	0.041	0.939	0.894	0.24
0.4532	20	N	1.000	0.036	-	-	-	0.789	0.697	-
		Y	0.989	0.041	1.210	1.318	0.108	0.779	0.698	-0.23
	40	N	1.000	0.019	-	-	-	0.886	0.832	-
		Y	0.994	0.019	1.268	1.335	0.067	0.880	0.822	-0.07
	80	N	1.000	0.010	-	-	-	0.939	0.906	-
		Y	0.997	0.012	1.302	1.382	0.080	0.938	0.908	-0.48
0.3707	20	N	1.000	0.033	-	-	-	0.799	0.539	-
		Y	0.994	0.056	1.170	1.333	0.163	0.787	0.532	0.20
	40	N	1.000	0.023	-	-	-	0.883	0.696	-
		Y	0.994	0.024	1.216	1.300	0.084	0.880	0.699	-0.35
	80	N	1.000	0.015	-	-	-	0.936	0.821	-
		Y	0.998	0.015	1.251	1.341	0.090	0.938	0.820	0.27
0.1336	20	N	1.006	0.134	-	-	-	0.799	0.405	-
		Y	0.998	0.126	1.134	1.383	0.249	0.795	0.409	0.23
	40	N	0.999	0.065	-	-	-	0.887	0.581	-
		Y	0.996	0.066	1.139	1.382	0.243	0.885	0.559	0.21
	80	N	0.999	0.039	-	-	-	0.939	0.737	-
		Y	0.996	0.043	1.287	1.345	0.058	0.936	0.742	-0.24
0.0774	20	N	1.005	0.119	-	-	-	0.776	0.232	-
		Y	0.989	0.166	1.335	1.403	0.068	0.803	0.175	0.02
	40	N	0.997	0.082	-	-	-	0.881	0.330	-
		Y	0.996	0.086	1.127	1.403	0.276	0.885	0.338	0.22
	80	N	0.998	0.059	-	-	-	0.939	0.480	-
		Y	0.996	0.057	1.124	1.397	0.273	0.942	0.490	-0.03

a Computed from all items when DIF=N, excludes the DIF item when DIF=Y
b Computed on DIF item only

Table 4

Experimental results for high difficulty *b* parameters (*1.0* to *2.0*) and *1:1* ratio of examinees, with observed score MH (MH_X), latent ability MH (MH_θ), MH_θ-MH_X (θ-X), test reliabilities for reference group (r_R) and focal group (r_F), and item for reference group (b_R).

% Overlap	#items	DIF	Ave MH_X[a]	Sd MH_X[a]	MH_X[b]	MH_θ[b]	(θ-X)	r_R	r_F	b_R
1.0000	20	N	1.000	0.000	-	-	-	0.685	0.685	-
		Y	0.986	0.001	1.273	1.381	0.108	0.710	0.709	1.07
	40	N	1.000	0.000	-	-	-	0.823	0.823	-
		Y	0.994	0.001	1.266	1.396	0.130	0.825	0.824	1.44
	80	N	1.000	0.000	-	-	-	0.906	0.906	-
		Y	0.996	0.000	1.312	1.386	0.074	0.906	0.906	1.01
0.8339	20	N	1.000	0.008	-	-	-	0.705	0.541	-
		Y	0.990	0.008	1.252	1.429	0.177	0.698	0.524	1.80
	40	N	1.000	0.004	-	-	-	0.831	0.706	-
		Y	0.995	0.005	1.268	1.438	0.170	0.831	0.705	1.69
	80	N	1.000	0.003	-	-	-	0.905	0.823	-
		Y	0.998	0.003	1.231	1.404	0.173	0.908	0.829	1.93
0.4532	20	N	0.997	0.061	-	-	-	0.711	0.420	-
		Y	0.995	0.064	1.091	1.332	0.241	0.697	0.385	1.70
	40	N	0.997	0.056	-	-	-	0.824	0.556	-
		Y	0.995	0.040	1.221	1.459	0.238	0.834	0.580	1.18
	80	N	0.998	0.031	-	-	-	0.906	0.736	-
		Y	0.996	0.031	1.200	1.366	0.166	0.904	0.729	1.52
0.3707	20	N	0.996	0.069	-	-	-	0.714	0.199	-
		Y	0.990	0.090	1.267	1.440	0.173	0.721	0.195	1.23
	40	N	0.998	0.056	-	-	-	0.833	0.315	-
		Y	0.999	0.058	1.026	1.348	0.322	0.833	0.335	1.92
	80	N	0.999	0.036	-	-	-	0.910	0.512	-
		Y	0.997	0.037	1.166	1.347	0.181	0.904	0.502	1.42
0.1336	20	N	1.013	0.141	-	-	-	0.714	0.084	-
		Y	1.004	0.199	1.011	1.320	0.309	0.703	0.069	1.85
	40	N	1.004	0.137	-	-	-	0.821	0.129	-
		Y	1.008	0.152	1.035	1.314	0.279	0.829	0.117	1.80
	80	N	1.001	0.120	-	-	-	0.906	0.250	-
		Y	0.999	0.119	1.317	1.380	0.063	0.907	0.227	1.09
0.0774	20	N	1.005	0.196	-	-	-	0.711	0.019	-
		Y	1.018	0.185	0.883	1.390	0.507	0.720	0.021	1.39
	40	N	1.008	0.169	-	-	-	0.837	0.042	-
		Y	1.000	0.158	1.341	1.442	0.101	0.829	0.049	1.01
	80	N	1.004	0.132	-	-	-	0.908	0.082	-
		Y	0.999	0.131	1.192	1.309	0.117	0.906	0.086	1.58

a Computed from all items when DIF=N, excludes the DIF item when DIF=Y
b Computed on DIF item only

Table 5

Experimental results for high difficulty *b* parameters (*1.0* to *2.0*) and *1:10* ratio of examinees, with observed score MH (MH_X), latent ability MH (MH_θ), MH_θ-MH_X (θ-X), test reliabilities for reference group (r_R) and focal group (r_F), and item difficulty for reference group (b_R).

%Overlap	#items	DIF	Ave MH_X[a]	Sd MH_X[a]	MH_X[b]	MH_θ[b]	(θ-X)	r_R	r_F	b_R
1.0000	20	N	1.000	0.000	-	-	-	0.709	0.709	-
		Y	0.989	0.001	1.321	1.464	0.143	0.678	0.675	1.69
	40	N	1.000	0.000	-	-	-	0.830	0.830	-
		Y	0.995	0.000	1.232	1.326	0.094	0.827	0.826	1.56
	80	N	1.000	0.000	-	-	-	0.906	0.906	-
		Y	0.997	0.000	1.248	1.346	0.098	0.906	0.905	1.44
0.8339	20	N	1.000	0.008	-	-	-	0.697	0.530	-
		Y	0.990	0.008	1.244	1.416	0.172	0.699	0.525	1.67
	40	N	1.000	0.005	-	-	-	0.830	0.704	-
		Y	0.996	0.004	1.208	1.320	0.112	0.825	0.696	1.83
	80	N	1.000	0.003	-	-	-	0.907	0.826	-
		Y	0.997	0.003	1.240	1.345	0.105	0.906	0.825	1.31
0.4532	20	N	0.999	0.066	-	-	-	0.700	0.393	-
		Y	0.995	0.061	1.111	1.317	0.206	0.711	0.403	1.66
	40	N	0.999	0.041	-	-	-	0.831	0.578	-
		Y	0.993	0.042	1.263	1.303	0.040	0.824	0.570	1.28
	80	N	0.998	0.034	-	-	-	0.907	0.735	-
		Y	0.996	0.030	1.162	1.296	0.134	0.906	0.734	1.55
0.3707	20	N	0.998	0.060	-	-	-	0.712	0.215	-
		Y	0.998	0.052	1.048	1.298	0.250	0.707	0.182	1.80
	40	N	0.998	0.057	-	-	-	0.827	0.318	-
		Y	0.996	0.057	1.136	1.322	0.186	0.820	0.328	1.88
	80	N	0.999	0.040	-	-	-	0.910	0.500	-
		Y	0.998	0.034	1.078	1.352	0.274	0.906	0.510	1.83
0.1336	20	N	0.995	0.166	-	-	-	0.705	0.080	-
		Y	1.001	0.174	1.065	1.348	0.283	0.688	0.060	1.70
	40	N	1.001	0.145	-	-	-	0.825	0.149	-
		Y	1.004	0.151	0.979	1.430	0.451	0.827	0.145	1.51
	80	N	1.000	0.108	-	-	-	0.906	0.250	-
		Y	1.001	0.102	1.131	1.334	0.203	0.906	0.251	1.55
0.0774	20	N	1.026	0.242	-	-	-	0.703	0.020	-
		Y	1.024	0.202	0.864	1.348	0.484	0.714	0.024	1.83
	40	N	1.012	0.148	-	-	-	0.835	0.043	-
		Y	1.001	0.149	1.260	1.370	0.110	0.833	0.038	1.31
	80	N	1.000	0.128	-	-	-	0.904	0.084	-
		Y	0.999	0.115	1.261	1.311	0.050	0.904	0.084	1.01

a Computed from all items when DIF=N, excludes the DIF item when DIF=Y
b Computed on DIF item only

Figure Captions

Figure 1. Difference in the MH common-odds ratios for moderate difficulty *b* parameters and *1:1* ratio. MH Difference is the absolute value of MH_θ-MH_X in the DIF-induced item.

Figure 2. Difference in the MH common-odds ratios for high difficulty *b* parameters and *1:1* ratio. MH Difference is the absolute value of MH_θ-MH_X in the DIF-induced item.

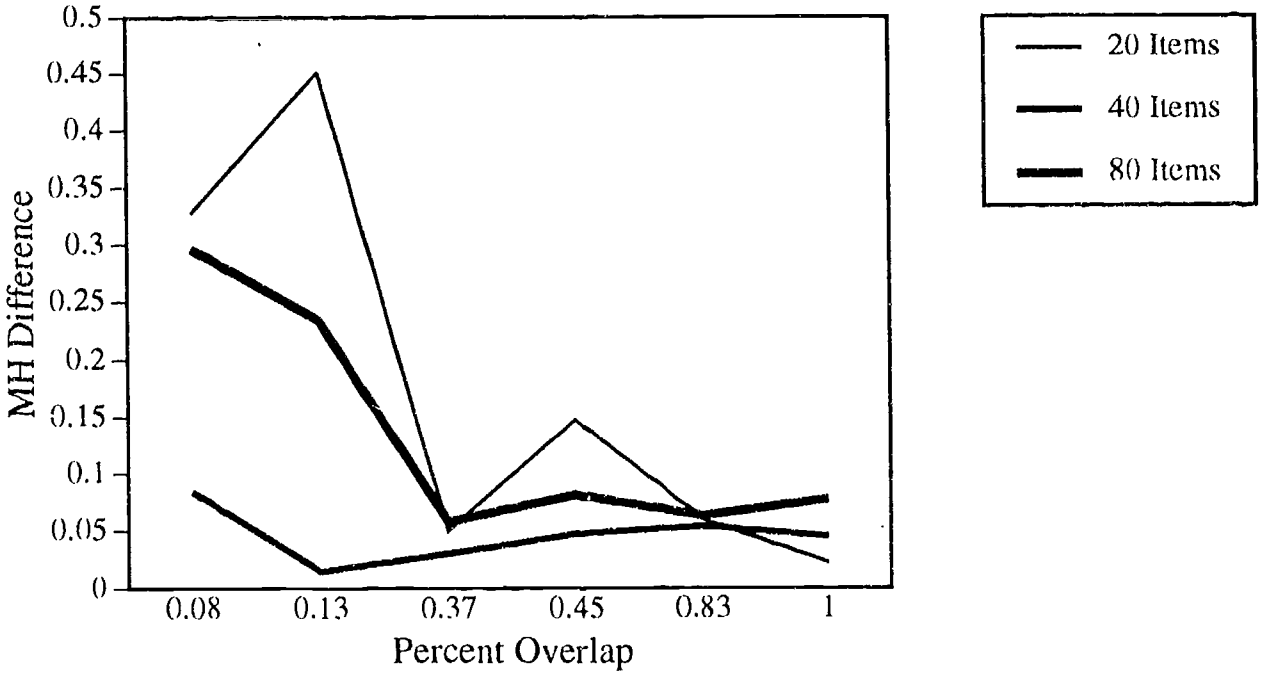
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
MH Difference
0.08
0.13
0.37
0.45
0.83
1
Percent Overlap
20 Items
40 Items
80 Items

Figure 1

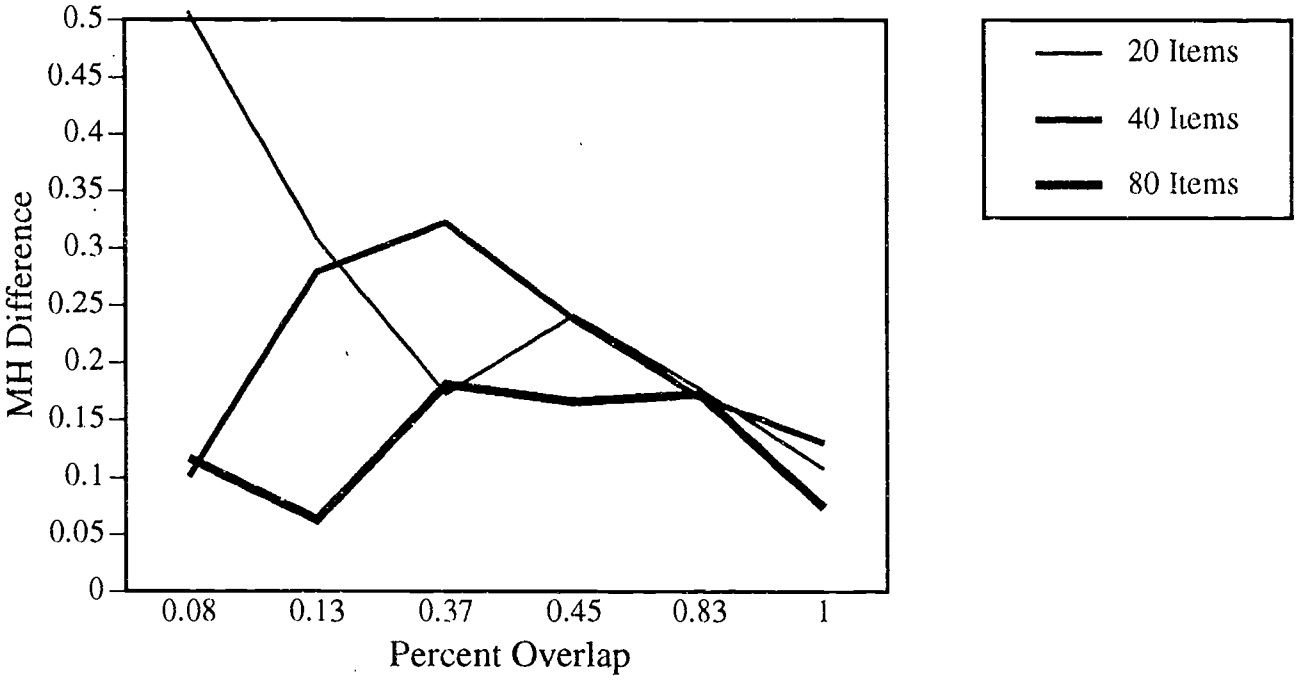
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
MH Difference
0.08
0.13
0.37
0.45
0.83
1
Percent Overlap
20 Items
40 Items
80 Items

Figure 2