NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-53251-3

Canada

CHARACTERISTICS OF THE DISTRIBUTION OF THE MANTEL-HAENSZEL DELTA UNDER DIFFERENT CONDITIONS OF THE NULL HYPOTHESIS: A MONTE CARLO STUDY

Thesis presented to the School of Graduate Studies and Research of the University of Ottawa in partial fulfillment of the requirements for the degree of Master of Arts in Education

Ottawa, Canada, 1989

ACKNOWLEDGEMENTS

Several people have earned my heartfelt gratitude for the help they gave me during the course of this study.

Dr. Marvin Boss, my supervisor, whose insights, suggestions, and encouragement never failed to provide me with the motivation necessary to continue my work.

Dr. Marc Gessaroli deserves my deepest thanks for sharing with me some of his great knowledge of data analysis and for allowing me to barge in on him everytime I encountered a problem.

I also owe a great debt to Len Fleming, senior consultant in computer programming at the University of Ottawa, who modified the data generation programs to fit my need. His availability to me was greatly appreciated and made my work infinitely easier.

Finally, I wish to thank my husband, Guillermo, and my sons, Alejandro, Rafael, and Nicolas. Without their love and support I never could have resumed my undergraduate studies, much less completed a Masters program. These very special people made it possible for me to indulge in "intellectual pursuits" while they were stuck with most (sometimes all) of the household chores. Thanks guys!

ABSTRACT

One of the concerns of educators today is to have bias-free tests. An increasingly popular technique to detect item bias or differential item functioning (dif) is the Mantel-Haenszel (M-H) delta. However, the effect of variables other than dif on this index has yet to be studied. The contribution of the present research lies in evaluating the effect of selected variables on the distribution of the M-H delta.

How would varying sample sizes, values of a common discrimination parameter ("a"), and values of the difficulty ("b") parameter affect the distribution of the M-H delta? The present research simulated data in order to evaluate the influence of these variables on the distribution of the M-H delta when no dif is present in the data.

Some confusion exists about which value of the M-H delta should be used to flag items as showing dif: both a value of "one" and of "two" have been proposed. What value(s) of the M-H delta would be appropriate to use as cutoff(s) with the data generated in this study? Do the variables mentioned above affect the cutoffs? The present research was designed to answer these questions.

Three sample sizes were used each with three to one ("majority"/"minority") ratios: 450/150, 900/300, and 1350/450. Three values of "a" were simulated: 0.7, 1.0, and 1.3. The "a" parameter was common to all items and to both "majority" and "minority" groups. A forty item test was simulated with each item having a specified difficulty ("b") parameter ranging from -2.0 to +2.0 with increments of .10. In order to generate a forty instead of a forty one item test, the "b" value equal to 0.0 was not used. One hundred replications were

generated for each combination of sample size and value of "a" conditions.

A 3x3 (sample size by value of "a") ANOVA was carried out on the standard deviation of the M-H delta computed across the 40 items. The two-way interaction was found to be significant at the .05 level. Increasing the sample size decreased the standard deviation of the index but increasing the "a" inflated the standard deviation; an exaggerated effect on the standard deviation was found when sample size was small (450/150) and "a" was large (1.3).

A repeated measures ANOVA was carried out on the standard deviation of the index across three groups of "b" values; that is, the first group included "b" values from -2.0 to -0.8, the second group included values from -0.7 to +0.6, and the third group of "b" values ranged from +0.7 to +2.0. The sample size by value of "a" by groups of "b" interaction was found to be significant (at alpha = 0.05) on the standard deviation of the index. The three-way interaction was explained as follows: as "a" increased, so did the standard deviation; when the "b" values were close to the center of the distribution, the standard deviation was smaller; this two-way interaction was exaggerated when samples were small and "a" large which contributed to the three-way interaction.

Four cutoffs were computed, two of which represented the .05 false positive rate, i.e. $P_{2.5}$ and $P_{97.5}$, and two corresponding to the .10 false positive rate, that is, P_5 and P_{95}. A MANOVA was carried out on these percentiles and the two-way interaction between sample size and "a" was found to be significant. The significant interaction was explained as

follows: as sample size increased, the cutoffs decreased; as "a" increased, the cutoffs were inflated; when sample sizes were small and "a" large, the effect was exaggerated, i.e. more than additive.

The proportions of false positive identifications (FPI) made when Wright's (1988) cutoff of "one" and Holland's (personal communication, April, 1987) suggested value of "two" seemed to be influenced by sample size and by "a". As sample size increased, the proportion of FPI decreased. As "a" increased, so did the proportion of FPI. Holland's cutoff of "two" was too stringent with the data studied here; that is, it always identified less than 5% of the items. Wright's cutoff of one, on the other hand, identified too many items (approximately 13%) when sample size was small (450/150). It identified approximately 5% of the items when samples were of moderate size (900/300) but flagged too few items (approximately 2%) with large samples (1350/450) except when "a" was large (1.3) where 4.2% of the items were flagged. Values of "b" close to the center of the distribution allowed less FPI to occur than when the values of "b" were located at either extreme of the distribution. Thus, with a false positive rate of .05 as the criterion, a cutoff of one would be appropriate to use only with moderate samples (900/300) or with larger samples when the "a" equals 0.7 or 1.0. In all other cases, Wright's cutoff of one and Holland's value of two would not produce a .05 false positive rate. This could then influence the proportion of false negative rates that the cutoffs would make if used when dif is present in the data.

Since all the variables studied here (sample size, value of "a", and value of "b") were found to have a significant effect on the standard

deviation of the M-H delta, it was concluded that the index is unstable across levels of these variables. The magnitude of the standard deviations found within conditions suggested that the index is also unstable across replications. A dif detection index should only be influenced by the amount of dif present in the data; in this study, the M-H delta was influenced by other variables. Therefore, there is cause for concern when this index is used.

Further studies could evaluate the effect of the variables studied here and/or others on the index when real data are used since the results of this study can be generalized only to similar data sets. Furthermore, studies in which dif is known should be done in order to evaluate the false negative rates that could be expected when different cutoffs are used.

The generalizability of this study should be proven by replications of its findings. Nevertheless, practitioners should be aware that variables other than the amount of dif may be influencing the M-H delta.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Social and political concerns have, for some time, influenced test developers to detect and correct any bias that could be found in their tests. Bias can exist in the test as a whole. A clear example of this is the use of a test in a selection situation where the performance of one group is not as well predicted as the performance of another group. In such a case, the selection criterion would include a discriminatory element against qualified members of the group whose behaviour was not well predicted. This type of bias is sometimes called "unfairness" and becomes evident when a test is used inadequately for a particular group.

A second type of bias is of interest in the present study and is found at the item level of the test. Such bias is not pervasive but rather specific to one or more items of the test. With this type of bias certain items of the test are found to be more difficult for one group than for another. Judgmental procedures have been developed in order to detect and correct or eliminate the biased items. Such procedures usually involve the examination of the items by one or more "expert judges" who are asked to determine if an item content or format appears more difficult for one group (usually the minority group) than for another group (usually the majority group) (Tittle, 1982). The judgmental method has been criticized for its heavy reliance on the subjectivity of the judges and is usually thought to be more effective if complemented by a more objective measure (Schmeiser, 1982; Plake, 1980).

Statistical methods developed to detect item bias are a more objective means of identifying problem items. These techniques cannot in and of themselves detect "bias" since a careful examination of the content and form of the flagged items would still be needed in order to determine if the problem is actually due to bias. As such, the items that are detected by the statistical procedures cannot be described as being biased. When statistical methods are used to flag items that perform unexpectedly differently, the word "bias" should then be replaced by the more appropriate term "differential item performance" (Welch, Ackerman, Doolittle, & Hurley ,1987; Wright, 1986), or by the term "differential item functioning" (Holland & Thayer, 1986). In this study, "differential item functioning" (dif) will be used.

Two very broad categories can be used to distinguish the statistical methods. These are: the unconditional and the conditional methods. With the unconditional methods the performance of two different groups on the item is compared. For example, the average performance of the Black examinees on an item could be compared to that of the White group and any significant difference would be considered evidence of dif. One disadvantage shared by all the unconditional procedures is their lack of information about the performance of examinees who share the same ability level but belong to different groups. In order to correct this disadvantage, the conditional methods were developed. The conditional methods use an estimate of the person's ability in order to compare the performance on the same item of persons with the same ability but belonging to different groups. Differences among such subgroups would be flagged as dif. Due to this definition of dif these

methods are described as conditional on ability.

One of the conditional methods, the Mantel-Haenszel procedure, has become increasingly popular. It was first developed by Mantel and Haenszel (1959) and later adapted by Holland (1985) for dif identification. The Mantel-Haenszel (M-H) procedure uses contingency tables in order to evaluate the performance of examinees with the same ability level but who belong to two different groups. The M-H approach yields different indices: a M-H chi-square, a M-H alpha, and a M-H delta. The indices are described in the next section.

Description of the Mantel-Haenszel Indices

The M-H Chi-square.

The Mantel-Haenszel chi-square was shown by Birch (1964) and Cox (1970) to be the uniformly most powerful unbiased test of the null hypothesis versus a specific alternative. This statistic is computed after organizing the data for each j^{th} score interval in a two by two contingency table such as:

		SCORE ON THE STUDIED ITEM		
		Correct	Incorrect	TOTAL
GROUP	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	n_{Fj}
	TOTAL	m_{1j}	m_{0j}	T_j

where the R refers to the reference (majority) group and the F is the focal (minority) group.

The hypothesis tested can be written as follows:

$$H_0 : \frac{P_{Rj}}{q_{Rj}} = \frac{P_{Fj}}{q_{Fj}} \qquad j = 1,\ldots,k$$

where p_{Rj} and p_{Fj} are the respective probabilities of success on the item of the reference (usually the majority) group and the focal (usually the minority) group at the j^{th} score interval; q_{Rj} and q_{Fj} are each group's respective probabilities of failure (1-p) for the same item at the j^{th} score interval.

The Mantel-Haenszel chi-square statistic is computed using the following formula:

$$MHCHISQ = \frac{(|\Sigma A_j - \Sigma E(A_j)| - 1/2)^2}{\Sigma Var(A_j)} \qquad (1)$$

where $E(A_j) = n_{Rj} m_{1j} / T_j$

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

The values of the elements in the preceding equations are taken from the contingency table shown above, and the summations are done across all ability levels. As can be seen from equation 1, the MHCHISQ will flag dif only when the dif is against (or in favor of) the same group along the ability scale. That is to say, the M-H procedure assumes that the dif is uniform. Non-uniform dif exists when, on the same item, one group is favored at some ability level but the other group is favored at a different level of ability. The M-H procedure will not detect non-uniform dif. A separate significance test should be used along with the M-H procedure to verify the absence of non-uniform dif. The MHCHISQ is an unsigned index since it cannot be negative. However, dif in either a positive or negative direction is maintained with this index up to the point where the sum of the differences is squared. As such this index although not showing the direction of dif by a

corresponding sign, does reflect dif in favor of one group or the other.

The MHCHISQ shares the same disadvantage as other chi-squares: large sample sizes inflate them; thus the MHCHISQ may flag items that do not necessarily have dif. Furthermore, in the study of dif what is of most interest is not the test of the null hypothesis but the evaluation of the importance of the difference between groups. Because of the sample size effect and the lack of information about the magnitude of the dif present in the data, another index is usually preferred over the MHCHISQ.

The M-H Delta.

The M-H delta is computed from the M-H alpha. This alpha is the common odds-ratio in the k 2x2 tables and its value is the odds ratio:

$$\text{M-H alpha} = \frac{\Sigma\, A_j\, D_j\, /T_j}{\Sigma\, B_j\, C_j\, /T_j} \qquad (2)$$

where the elements are taken from the table shown above and the summations are done across all ability levels.

A value of alpha equal to one can be seen to mean there is no difference between the two groups. A M-H delta can be computed by

$$\text{M-H delta} = -2.35\, \ln(\text{M-H alpha}) \qquad (3)$$

and a resulting delta value of zero is easily interpreted as corresponding to the absence of dif. The delta used as a measure of item difficulty by ETS has a mean of 13 and a standard deviation of four. The distribution of the M-H delta should have a mean of zero but its standard deviation is not known.

Both these values (alpha and delta) have quite simple interpretations. The alpha is the average factor by which the odds are greater that an examinee from the reference group succeeds on the item than a member of the focal group with the same ability. When the alpha value is greater than one, the members of the reference group performed better on average on this item than did the members of the focal group with the same ability levels. The delta values, on the other hand, denote the difference in difficulty of an item for the examinees of the reference group compared to members of the focal group within the same score intervals (ability levels). Wright (1986) notes that a delta value of one is usually considered important, although Holland (personal communication, April, 1987) suggests that a value of two would be more appropriate. Holland and Thayer (1986) have reported that the M-H alpha is a consistent and accurate estimator in the range $1/3 < \text{M-H alpha} < 3$ (delta values between -2.6 and +2.6).

Holland and Thayer (1986) have determined that under certain conditions, the M-H yields identical results to the one-parameter logistic model. The conditions are (a) that the samples be representative of the subpopulations studied, (b) that the total test score not include the items flagged in a preliminary analysis as exhibiting dif and (c) that the studied item always be included in the total test score, regardless of the results of the preliminary analysis. For example, if the first analysis identified five items as exhibiting dif, then the subsequent analysis would eliminate these items from the total score when other items were studied but when one of these five items was being analyzed it would be included in the

total score. The authors propose that such conditions be present when the M-H procedure is used. Since the consequences of using these criteria have not been systematically studied, the authors consider them only tentative improvements of the Mantel-Haenszel technique.

According to Holland and Thayer (1986) if the three conditions mentioned above are met, the M-H alpha takes the form

$$\text{M-H alpha} = e^{(b-b')} \qquad (4)$$

where b and b' are the difficulty parameters for the focal and the reference groups, respectively.

Equation 4 implies that the discrimination parameter ("a") is equal to one in the following equation

$$\text{M-H alpha} = e^{a(b-b')} \qquad (5)$$

However, this is a double assumption for the one-parameter logistic model (i.e. common "a" and value of one) that is more difficult to accept. Allen and Yen (1979) describe the one-parameter logistic model as assuming a common "a" but without also assuming a value of one. The effect of a common "a" that would differ from one is obvious from equation 5 when a difference between "b's" exists. This effect has yet to be reported in studies evaluating the performance of the M-H alpha (or delta). Furthermore, if there is no difference between the "b's" and if equation 5 holds, then the M-H alpha should always be equal to one (and the delta equal to zero) regardless of the value of the "a". The effect of different values of a common "a" has also yet to be studied under the null hypothesis. Notice that the common "a" whether equal to one or not is assumed to be equal for both groups. If it were not, non-uniform <u>dif</u> would exist and the M-H procedure may not detect

it.

One clear disadvantage of the M-H alpha is its lack of an accurate error variance; only approximations are known. Of these approximate terms (i.e. Breslow, 1981; Flanders, 1985; Hauck, 1979; and Phillips & Holland, 1986), the estimates of Breslow and of Flanders seem to be the most accurate (Linacre, 1988).

Various studies have been conducted in which the researchers evaluated the M-H procedure against some other dif detection technique. These studies give some information about how the M-H indices perform. In the next section, the studies are described in order to evaluate the amount and the scope of the information available about the M-H procedure.

Studies of the M-H Indices

A number of researchers, using the M-H procedure, examine the possibility of dif in a particular test. In such studies dif is expected in the test but its amount is unknown. Three examples of such studies will be mentioned here.

Martois, Rickard, and Stiles (1988) examined the Greater Avenues to Independence (GAIN) test to see if Anglo, Hispanic, and Black examinees showed more dif on the reading items of the test compared to the mathematical section of the GAIN (dif, favoring the Anglo group, was expected on the reading items but not on the math subtest). Miller, Doolittle, and Ackerman (1988) compared the performance of ESL students to that of non-ESL examinees on an English usage and a mathematics test (with dif expected to favor the non-ESL group on the English but not the math test). A final example can be found in the study by Welch and

Doolittle (1988) who used the M-H procedure to evaluate the performance of males and females on English usage items. Females were expected to perform differentially better on grammar, punctuation, and sentence structure items, and males were expected to perform better on logic-organization and diction-style items.

All of these researchers although expecting one group to perform better on certain subtests do not have knowledge of which items should be flagged as showing dif and to what extent those items favor one group. As such, this type of study cannot adequately evaluate the performance of a dif index per se and does not provide information on the meaningfulness of the results obtained by the procedure. Nevertheless, if the expected differences were found it would suggest that the distributions of dif reflect what was hypothesized and consequently that the procedure adequately identifies dif. However, none of these studies found the expected differences.

Studies comparing the performance of the M-H procedure to other dif detection techniques may lead to a better understanding of how the M-H indices function. Several of these comparative studies are reviewed here.

Dizinno and Arrasmith (1988) compared the results of a subjective method of bias detection (item review) to dif as identified by the M-H delta. They used data from the spring 1987 administration of the New Mexico High School Proficiency Exam. This test includes 100 items assessing life skills in five major areas (community resources, consumer economics, government and law, mental and physical health, and occupational knowledge). From the original population, only Anglo

Americans and Native Americans were chosen since they differed significantly on their passing rates (97% and 65%, respectively). These two groups were then matched on ability by randomly selecting Anglos at each score point of the raw score distribution of the Native Americans. This restricted the range of scores from 51% to 85%. There were 1,518 examinees in each of these two groups. Based on the M-H delta estimates, three groups of items were chosen to be reviewed. The first group consisted of items with a M-H delta greater than one, i.e. with dif against the Anglo group (13 items). The second set of items had M-H delta values more extreme than minus one and were exhibiting dif against the Native group (12 items). Finally, the M-H delta values for the third group were in the range of plus to minus .20 and were not considered biased (14 items). The authors also studied the "impact", as opposed to the dif, of these 39 items. Impact is defined as an item for which one group, as a whole, performs differently than another group. Dif, as will be recalled, is defined as differential performance of examinees with the same ability who belong to different groups. The author suggested evaluating the impact of the items by comparing the p-values of the two groups (before matching) without dividing the data into score categories. If the difference between the two groups in the difficulty (p-value) of an item was greater than .1, then the authors concluded there was differential impact of the item.

The 39 chosen items were reviewed by 90 tenth-grade students and by 29 faculty members. The students answered two questions about each item, one concerning their ease in answering the item and one about how sure they were of having responded correctly. The teachers, on the

other hand, completed an item bias review form which consisted of 14 questions about the content of the item, its wording, the answer choices, and the degree of difference of the two groups concerned (Anglos and Natives) in their opportunity to learn.

The students' reviews were disappointing when compared to the items flagged by the M-H delta since their judgments were often contrary to the findings using the index. The teachers' reviews, on the other hand, were more related to differences in the difficulty level of the item for the two groups (impact) than to the value of the M-H delta. On the whole, the researchers did not find a relationship between the performance of the M-H delta and any possible "bias" as detected by the reviewers. The relationship between p-value differences and the M-H delta was not studied. The results of this study suggest that the M-H delta may not identify the same items as would be detected in an item review. The relationship between a subjective classification of dif and the results obtained when statistical techniques are used was also found to be small if not completely lacking in a number of other studies (e.g. Ironson & Craig, 1982; Phillips & Mehrens, 1988; Simon, 1987).

Hambleton and Rogers (1988) compared the performance of the M-H chi-square (MHCHISQ) to that of Rudner's (1977) unsigned area index (UA). The UA is based on a three-parameter logistic model and defines dif as a difference between the item characteristic curves (ICCs) of the two groups. The UA is given by

$$UA = \Sigma \left[\left| P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \right| \right].005$$

where $P(u_g=1|\theta_i)$ and $P'(u_g=1|\theta_i)$ are the probabilities of a correct response for the majority and the minority group respectively and the summation is done across ability levels, from -5.0 to +5.0.

The authors compared the MHCHISQ to the UA in order to determine if the MHCHISQ is similar enough to the UA to be substituted for it, and to examine the behaviour of these statistics when the ability distributions of the two groups are considerably different.

Hambleton and Rogers used the items of the New Mexico High School Proficiency Exam assessing life skills that were described by Dizinno and Arrasmith (1988). Only 75 items of the test were used excluding very easy items ($p > .90$) and items with very low discrimination ($r_{bis} < .1$). The test had been administered to 8,000 Anglo-Americans and 2,600 Native Americans. The data set was chosen because of widely discrepant score distributions of the two groups. Two equivalent comparisons were carried out in order to evaluate the consistency of the dif classification of each statistic. Two equivalent Anglo (A1 and A2) and two equivalent Native samples (N1 and N2) were drawn from the original population. Each of these subsamples contained 1,000 examinees. The three-parameter logistic model was fitted separately to the four subsamples.

The dif classification using subsamples A1 and N1 was compared to that using subsamples A2 and N2. The consistency of each statistic was evaluated by computing the percentage of items classified consistently (as dif or no dif) in the two independent comparisons. Since the distribution of the area statistic is not known, a cutoff score was obtained in this study from a comparison of two randomly equivalent

groups (Native Americans). The highest area statistic value thus obtained served as the cutoff value in the subsequent analyses since it represented the highest value occurring by chance.

The M-H chi-square was computed in a two step procedure. First, the score groups were formed based on all items, the statistic calculated, and the items flagged if the chi-square exceeded the value at the .01 level of significance. Then "purified" total scores, excluding the flagged items, were recalculated, the score groups reformed, and the MHCHISQ recomputed. Notice that the authors do not seem to have used the Holland and Thayer (1986) criterion of including any item in the purified total score when that item was studied regardless of the results of the preliminary analysis.

Both statistics showed moderate consistency across equivalent comparisons, the UA being the least consistent (73% of consistently classified items) and the M-H chi-square the most consistent (80%). Since the amount of dif present in the test is unknown, the percentage of agreement that could be expected by chance cannot be computed. However, the consistent classification by each statistic can be compared to the total number of items in order to determine which index consistently classified the greatest proportion of items. Note that since the indices are both unsigned, the direction of dif was not studied.

The UA consistently flagged 55 out of a possible 75 items (14 dif and 41 no dif) while the MHCHISQ consistently flagged 60 of the 75 items (9 dif and 51 no dif). Some items consistently identified as exhibiting dif by the MHCHISQ can be considered a subset of the ones

flagged by the UA (14 items flagged by the UA, 9 by the M-H of which 7 are common).

Nine items were consistently flagged by one method but not by the other. When, out of four analyses (two indices computed on two independent comparisons), an item was flagged in all but one instance, the authors considered that instance to be due to a type II error. Such an item would have been flagged by one index in both comparisons but only once when the other index was used. In such a case, the dif was considered "real" and the lack of complete agreement between the indices indicative of a type II error, i.e. failure to detect a difference that exists.

Using this criterion, the UA index made one type II error and the MHCHISQ made two. Six items were consistently flagged by the area index but not at all by the M-H. Of these items, five plotted ICCs crossed markedly which suggests non-uniform bias that the MHCHISQ does not flag (a significance test does exist which may detect this interaction but was not used in this study). The sixth item flagged by the UA and not by the MHCHISQ suggests that the latter index may be sensitive to the number of examinees per group per score category. For this item, few (about 15%) Anglos scored in the region of the scale where the largest difference was observed. The UA did flag such an item because the ICCs were markedly different.

Another example of the possible sensitivity to the number of subjects per group per score category of the MHCHISQ was found in one item flagged by the MHCHISQ but not by the UA; plotting of the ICCs showed uniform but not marked differences. The most marked difference

between the ICCs was in the range where many Natives scored which probably is why the MHCHISQ did flag it. The area index did not identify this item as exhibiting dif because the ICCs were not markedly different. However, some caution should be exercised when interpreting these results since the UA may also be sensitive to the number of examinees per group per score category. Few examinees at any given ability level will lead to poor estimates of the parameters of items appropriate for that ability which in turn could be responsible for the dif classification.

In order to study the effect of differing distributions on the area statistic, the ability continuum analyzed was changed to a range from plus to minus two standard deviations from the mean of the Native group, focusing the analysis on the part of the ability scale where most of the focal group lies and therefore where the largest practical difference between the two groups might be found. A matched group analysis was conducted in order to study the effect of the original differing distributions on the M-H chi-square: one third of the Native group was selected to closely match the distribution of the Anglos (N=650). Both the UA and the MHCHISQ were computed with the matched groups.

With the matched sample analysis, moderate change was noted in the MHCHISQ (all nine previously flagged items still identified and four additional ones flagged). A great deal of difference was observed in the UA which is probably due to the smaller sample size (only five of 14 previously identified items were flagged).

The two indices were found to be somewhat inconsistent when independent equivalent comparisons were carried out. This reinforces the belief that flagged items should only be treated as "potentially biased". There is quite a good degree of agreement between these two methods as to the identification of uniform bias; the UA seems to detect non-uniform bias whereas the MHCHISQ does not; however, as noted above, a test does exist to examine the presence of such an interaction but it has not been adequately tested yet and was not used in this study. The ability distributions of the two groups do not seem to have any major impact on the MHCHISQ although matching does somewhat alter the results. More change was noted in the UA with the matched analysis which may be due more to the decrease in sample size than to differences in the distributions. The MHCHISQ appears to be more stable with small sample sizes. The authors suggest that the MHCHISQ could be safely substituted for the UA if safeguards are put in place to detect non-uniform bias.

Skaggs and Lissitz (1988) evaluated the consistency of the MHCHISQ, the M-H delta, and five other indices which are briefly described here.

1. Rudner's (1977) unsigned area index (UA)

   This index was described above.

2. Lord's unsigned chi-square

   Lord's (1980) chi-square simultaneously compares differences in the discrimination ("a") and the difficulty ("b") parameters. This test is computed for each item and approximates the chi-square distribution with two degrees of freedom. It is given by

$$LORD = (V_{i.j=1} - V_{i.j=2})' (\Sigma_{i.j=1} + \Sigma_{i.j=2})^{-1} (V_{i.j=1} - V_{i.j=2})$$

where V_{ij} =the matrix of a and b parameters for subgroup j;

Σ_{ij} =the covariance matrices for the vectors of "a" and "b" parameters for subgroup j.

3. The SOS2 and the SOS4

The SOS2 and the SOS4 indices compute the sum of squared differences between groups of the probabilities of obtaining a correct response on an item at each level of ability. These indices are weighted (by the inverse of the error variance of the difference in ICCs at each ability level) thus taking the estimation error into account. The SOS4 is a signed index while the SOS2 is unsigned and they are computed as follows:

$$SOS2 = \frac{1}{n_R+n_f} \sum \frac{\{P_{iR}(\Theta_j) - P_{iF}(\Theta_j)\}^2}{\sigma^2_{PiR-PiF}}$$

$$SOS4 = \frac{1}{n_R+n_f} \sum \frac{\{P_{iR}(\Theta_j) - P_{iF}(\Theta_j)\}\,|P_{iR}(\Theta_j) - P_{iF}(\Theta_j)|}{\sigma^2_{PiR-PiF}}$$

where $P_{iR}(\Theta)$ and $P_{iF}(\Theta)$ are the estimated probabilities of a correct answer for the reference group and the focal group respectively;

the j subscript refers to all instances of theta for either group $(n_R + n_f)$;

Θ_j is an ability level observed in the reference (or the focal) group and the probability difference computed as if that level of ability was also observed in the focal (or reference) group;

n_R and n_F are the number of examinees in the reference group and the focal group respectively;

and $\sigma^2_{PIR-PIF}$ is the variance error of the difference in estimated probabilities.

The summations are done across all ability levels.

4. Camilli's chi-square

Camilli's (1977) chi-square is computed as a conventional chi-square for every 2x2 contingency table and then summed across intervals of ability level. The unsigned version of this index is given by

$$UCAM = \Sigma \frac{n_{1j}\, n_{2j}\, (P_{1j} - P_{2j})^2}{(n_{1j} + n_{2j})\, P_{\cdot j}(1 - P_{\cdot j})}$$

where P_{1j} and P_{2j} are the proportions of examinees in group one and two, respectively, who scored in the j^{th} category and answered the item correctly.

$P_{\cdot j}$ is the proportion of subjects from both groups who responded correctly and scored in the j^{th} category.

The summation is done across all intervals of ability.

Two forms of a basic skill mathematics test were administered, and 96 items were analyzed and evaluated for differential performance of males and females. The examinees, eight graders, were divided into three sampling conditions: (1) a 1986 sample consisting of 650 females and 650 males; (2) a 1987 sample of 2,000 females and 2,000 males; and (3) a reduced 1987 stratified sample of 650 females and 650 males. Males and females had virtually identical raw score distributions hence no overall differences in achievement level.

The majority of the correlations of an index with itself across

samples were under .70 which is quite low for consistency if dif were really present. A number of items are assumed to have no dif and if no dif existed in reality, the correlations could be due to chance occurrence. Within sampling conditions, the intercorrelations were high between the SOS4 and the M-H delta (from .84 to .90) which is not surprising since both of these indices are signed. The next largest correlations were between the UA and the ULORD (.80 to .88) and between the MHCHISQ and the UCAM (.57 to .84). The rest of the intercorrelations among the unsigned indices were modest (in the .4 to .6 range). The correlations between the signed and the unsigned indices were close to zero.

In order to obtain a baseline from which to evaluate the indices, two randomly equivalent samples (2,000 each and for each of which the ratio of males to females was equal) were formed from the 1987 sample (not reduced). An index value was defined as P_{80}, that is, the value separating the upper 20% of the distribution of each index from the remainder of the distribution and was the arbitrary criterion used to classify dif items. This arbitrary criterion was chosen in order that the statistics flag the same number of items thus reducing the effect of the number of dif and no dif items on the percent of agreement. (The actual values used as cutoffs were not provided in this study). Most percentages of agreement in dif classification were in the 70's with more agreement among techniques when only one sampling condition was studied and less agreement when different sampling conditions were considered. A percentage agreement of 70 is not a very large improvement over chance since with a proportion of .8 no dif and .2 dif

items 68% agreement by chance would be expected. The observed low agreement could be explained in part by the lack of consistency of the indices or by the fact that perhaps there were actually few dif items in the test and largely only chance occurrence was being examined. Another possible explanation for the indices' low consistency is that other variables could mask the sex differences, especially since the 1986 sample was not randomly selected but had been chosen for a pilot study and as such is not comparable to the 1987 samples.

Six replications of the IRT analyses were done (i.e. two- and three-parameter model estimations by three sampling conditions). The UA did not flag any item in all six cases and only three items were identified in five of the six replications. LORD, SOS2, and SOS4 were more consistent flagging, respectively, the same one, four, and five items in each of the six replications.

Three replications were done with the non-IRT methods (i.e. the three sampling conditions). The MHCHISQ and M-H delta flagged, respectively, six and five items in all three cases while UCAM identified four items. An examination of the content of the items flagged by all seven indices did not reveal any reasons for the dif classification. The consistency of the indices is further questioned by their failure to identify the same dif items in samples from different years. But once again it must be stated that since the amount of dif was unknown, the indices' performance could be a measure of random occurrence rather than evidence of any inconsistency.

Welch, Ackerman, Doolittle, and Hurley (1987) used the American College Testing Assessment Test (Form 26E) in order to evaluate the

performance of American compared to that of Irish examinees. They analyzed the items relating to English Usage and to Mathematics Usage. The English Usage subtest consists of 75 multiple choice items measuring punctuation, grammar, sentence structure, diction and style, and logic and organization. The second subtest is a 40-item multiple choice mathematical reasoning ability test. The sample consisted of 2,738 college (university) bound Americans and 2,499 Irish applicants to a technological university in Ireland. Various dif indices were compared to the MHCHISQ which was supplemented by a log-linear test to verify the absence of non-uniform dif. The other indices used in this study are described here.

1. The delta-plot method (TID)

Angoff and Ford's (1973) delta-plot method, also known as transformed item difficulty (TID) is an unconditional technique which evaluates the differences in difficulty levels (p-values) for the two groups. The p-values are calculated separately for each group. These values are then transformed to normal deviates (z-scores) which are in turn transformed to delta values (delta= 4z+13). The delta values for both groups are then plotted against each other and should form an ellipse around the major axis defined as

$$Y = AX + B$$

where Y = the delta value for one group

X = the delta value for the other group

$$A = \frac{(s_y^2 - s_x^2) \pm [(s_y^2 - s_x^2)^2 + 4r_{xy}^2 s_x^2 s_y^2]^{.5}}{2r_{xy} s_x s_y}$$

and $B = M_y - AM_x$

The values M_x and s_x are the mean and standard deviation of the deltas for the group plotted on the x-axis. The M_y and s_y are the corresponding values for the group plotted on the y-axis. Finally, r_{xy} is the correlation between the deltas for the two groups.

Angoff and Ford operationally describe dif as the deltas' distance to the major axis of the plot. Any outlying item is considered problematic and thus flagged. The perpendicular distance index is computed as follows for each item:

$$D_i = \frac{AX_i - Y_i + B}{(A^2 + 1)^{.5}}$$

2. A modified TID

A modified version of the delta-plot method was also used in this study. This index (TID-dis) partials out from the delta indices the variance accounted for by the point-biserial correlations (of item with total score) for the combined group.

3. Linn and Harnisch's index

Linn and Harnisch (1981) developed a lack of fit index which is known as the pseudo-IRT or the modified caution index (MCI). With this procedure, the combined groups are used to estimate the "a", "b", and the "c" parameters and the three-parameter model is used to determine the expected probability that an examinee would answer an item correctly. The difference between the expected and the observed probabilities for the focal group are computed and averaged for various previously determined ability intervals. The MCI index is the sum across intervals of the standardized differences. For each ability level, the average of the standardized differences is as follows:

$$MCI_{iq} = 1/n_q \{\Sigma(U_{ij} - P_{ij} / [P_{ij}(1-P_{ij})]^{1/2})\}$$

where U_{ij} = 1 if correct answer, 0 if incorrect

q = number of intervals on the ability scale

n_q = number of individuals in the q interval

and P_{ij} = probability of a correct response for the combined subgroups.

Summing across intervals, the overall standardized difference for the minority group is

$$MCI_i = \Sigma (n_q MCI_{iq}) / \Sigma(n_q)$$

4. SOS3

The SOS3 is a sum of squares index similar to the SOS4 described above except that the SOS3 is unweighted and is computed as follows:

$$SOS3 = \frac{1}{n_R + n_F} \Sigma (P_{iR}(\Theta_j) - P_{iF}(\Theta_j)) \left| P_{iR}(\Theta_j) - P_{iF}(\Theta_j) \right|$$

where the elements of the equation are the same as those described for the SOS4.

Three other seldom used dif indices were also computed: two ordinary least squares regressions, one of the Irish (p_I) onto the difficulty of the Americans (p_A) and another least squares regression of the arcsine-transformed difficulty value of the p_I onto p_A (the residuals were examined for outliers); and a principal component analysis where the second component scores were then examined for outliers.

The Spearman rank order correlations among the indices were all substantial (from .82 to .99) except for the SOS3 index (from .50 to .67 in the English test and .69 with the TID-dis in the math test). Overall, the intercorrelations were highest on the math test. The

MHCHISQ correlated well (>.90) with all the indices except the TID-dis (.82) on the math test and the SOS3 (.67) on the English test.

Wright (1986) compared the performance of the M-H delta to that of three indices of the standardization approach . The standardization method, described by Dorans and Kulick (1986), uses a two by two by score interval table with equal number of subjects at each interval and yields three indices.

The root mean weighted square difference is computed as follows:

$$RMWSD = [\Sigma K_s (P_{fs} - P_{Bs})^2 / \Sigma K_s]^{.5}$$

where K_s is the relative frequency at each score level s of the focal group;

P_{fs} and P_{Bs} elements refer to the proportion correct responses by the focal and the base groups respectively at total score level s.

The authors describe another index which can be used to detect dif: the standardized p-difference (STD-D) which is defined as

$$STD\text{-}D = \Sigma K_s (P_{fs} - P_{Bs}) / \Sigma K_s$$

A log transformation (proposed by Holland, 1985) provides a statistic on the delta scale that is computed as follows:

$$STD\text{-}delta = -2.35 \ln \frac{p'_b / (1-p'_b)}{p'_f / (1-p'_f)}$$

where $p'_f = \Sigma w_s p_{fs}$

$p'_b = \Sigma w_s p_{bs}$

w_s = relative frequency at total score level s of the standardization group (usually the focal group)

p_{fs} = probability of correct response for the focal group members in category s

p_{bs} = probability of correct response for the base group members in category s

For both the M-H delta and the STD-delta Wright used a criterion of one to identify dif.

The data were gathered from a 1984 administration of the Scholastic Aptitude Test. From the original population, 10,000 Whites and 3,000 Blacks were drawn and represented the full sample in this study. Subsamples of 3,000 Whites/ 3,000 Blacks, 10,000 Whites/800 Blacks, 5,000 Whites/400 Blacks, 2,500 Whites/200 Blacks, and 1,000 Whites/ 80 Blacks were constructed in order to study the performance of the indices under different sample size conditions. Three replications were performed on each of the last three sample sizes in order to evaluate the consistency of the indices within different sample size conditions.

The effect of the sample size was evaluated first by the mean square difference (MSD) between values of the statistic in the full population and one of the other samples. The MSD is given by

$$MSD = \Sigma (x_{if} - x_{ij})^2 / n_i$$

where x_{if} = the value of the statistic "x" for item i in the full sample

x_{ij} = the value of the statistic "x" for item i in sample j.

n_i = the number of items.

The second statistic used to evaluate the consistency of the indices across sample sizes was the correlation between the same index in the full sample and in each subsample.

As expected, the MSD increased as the sample size decreased. For the smallest samples, all methods showed correlations between the same index in the full sample and in each of the subsamples in the .5 to .6 region. This does not seem adequate for a routine item screening procedure.

The score intervals were also manipulated analyzing the data with 61 ten-point and with six one-hundred-point intervals. Pearson correlations between the M-H delta and the STD-delta within sample sizes and score intervals were high (from .94 to .99). The number of score intervals had little effect on the ordering of the items (according to their index value), but it had an important effect on the actual size of the indices with their value being consistently lower when fewer intervals were used. The need for more intervals may be due to the substantial difference between the mean ability of these two groups.

The results obtained when the RMWSD was used are not detailed here because it was shown that this index retains the sampling error. For example, the mean value for this index was 0.077 with the full sample (10,000 Whites, 3,000 Blacks) but its mean value increased dramatically when smaller sample sizes were used: from .15 to .18 with the three replications of 5,000 Whites/400 Blacks; the mean RMWSD was approximately .20 with 2,500 Whites and 200 Blacks; and its value climbed to around .29 when the sample consisted of 1,000 Whites and 80

Blacks. Also, Pearson correlations across replications were in the .2 to .6 range. Finally, although increasing the interval size for the RMWSD did increase the correlations with the values obtained from the full sample, these coefficients were still lower than the ones computed with the other indices.

Phillips and Mehrens (1988) used both simulated and real data to compare the performance of five indices: the Linn and Harnisch index (MCI); Angoff's delta plot (TID) with matched groups using the major axis and its modified version using the 45 degree line (TID-45); differences (t-tests) in the "b" parameter of the one-parameter logistic model (ICC-1_b); a modification of the ICC-1_b where the no dif items are adjusted to a mean difficulty of zero in each group (adjusted ICC-1_b); and the Mantel-Haenszel chi-square (MHCHISQ). These indices had explicit criteria for dif: a MCI value above 2.6 or below -2.6; a TID or a TID-45 perpendicular distance to the major axis of more than one standard deviation; the difference in the "b" estimates exceeding the sum of the estimated standard errors for an item in the two groups (i.e. confidence intervals for the two item "b's" which do not overlap); and the M-H chi-square critical value at the .05 level of significance.

The real data came from sixth-grade math students in one school. The Stanford Achievement Test was administered to all students in the spring. The opportunity to learn was evaluated by classifying the assigned exercises for the year into a math content matrix. The items in the Stanford Achievement Test were also classified using the same matrix. The first group (with opportunity to learn) consisted of 69

students from three teachers, the second group (without opportunity to learn) included 87 students from three teachers. The test items to be analyzed were chosen according to the matrix classification. Three sets of items were studied: (a) items with good and equal coverage for the two groups (no dif); (b) items with no or very little coverage for both groups (no dif); (c) items with good coverage for one group but little or no coverage for the other group (dif). In all, there were 14 no dif and 17 dif items.

The various indices did not perform very well. The TID with matched groups correctly identified 10/14 no dif items and 6/17 dif items and three of the true dif items were flagged in the wrong direction. Using the TID-45 index with matched groups, all the no dif items and 7/17 dif items were correctly identified. The MCI index correctly classified only one dif item. Finally, the MHCHISQ correctly identified 13/14 no dif and 6/17 dif items. The indices with the largest percentages of correct dif classification were the TID-45, the unadjusted ICC-1_b, and the adjusted ICC-1_b (72%, 74% and 86%, respectively). The TID, the MCI index and the M-H chi-square had quite low percentages of correct decisions (52%, 48%, 61%, respectively). Since many items were not correctly classified, these results are related to the findings of Dizinno and Arrasmith (1988) described above who stated that dif techniques do not flag the same items as are identified with a subjective procedure.

The authors also compared indices computed from simulated data. The simulated data, generated with the one-parameter logistic model, contained two sets of 500 thetas each (estimated without dif). The

bias was manipulated with increments on the "b" parameter estimates for one group. Four dif conditions resulted: (a) 20 items with their original "b" estimates (no dif); (b) 10 items with a small amount of dif (0.1 logit increase); (c) 10 items with moderate dif (0.5 logit increase); and (d) 10 items with large dif (1.0 logit increase).

With these data, the ICC-1_b incorrectly identified a large proportion (14/20) of no dif items as exhibiting dif against group one (with opportunity to learn) but the adjusted ICC-1_b index did better, correctly identifying 12/20 no dif items, five out of ten items with a small amount of dif, and all the items where medium and large amounts of dif were induced. The TID correctly classified 12/20 no dif items, none of the items with small or medium dif, and nine of the ten large dif items. The TID-45 correctly identified all the no dif items but none of the items with a small amount of dif; six moderate dif items were flagged along with all the large dif items. The MCI index identified five no dif items as exhibiting dif, no items with small or medium amount of dif, and seven out of the ten items with the largest implanted dif. The MHCHISQ flagged 13 no dif items as showing dif against group one (with opportunity to learn), small and medium size dif went almost undetected (0/10 and 1/10, respectively), but all the items where large dif was induced were flagged. The percentages of correct classifications into dif and no dif items were as follows: the ICC-1_b, 44%; the adjusted ICC-1_b, 56%; the TID, 42%; the TID-45, 72%; the MCI index, 44%; and the MHCHISQ, 36%. With a 40%/60% no dif/dif split, 52% of agreement would be expected by chance (i.e. $4^2 + 6^2$). After changing the proportion of no dif to dif items from the original

40%/60% to a new 80%/20% split, only the MHCHISQ was computed (since it did the least well with the first split). As with the 40%/60% split, approximately equal numbers of dif and no dif items were flagged by the MHCHISQ. This index was still somewhat insensitive to small and medium bias but the percentage of correct classification increased to close to 80% (68% would be expected by chance with this new split).

The results of this study must be interpreted cautiously for various reasons. It must be remembered that since the simulated data were generated with the one-parameter model, the ICC-1_b, whether adjusted or not, could be favored over the others. The M-H procedure was suggested by Holland and Thayer (1986) to give similar results as the one-parameter index only under certain conditions which were not followed here. Another problem facing a fair interpretation of these results is that the 40%/60% split is quite high since one of the assumptions for using dif indices is that the bias is not pervasive. The performance of the MHCHISQ was much better when the 80%/20% split was used. Also, the very small sample sizes used in the real data (n_1= 69; n_2= 87) cannot be expected to yield very stable results and must put in question the techniques' success in flagging the dif in the real data.

The studies described above compared one or both M-H indices to different techniques and used different methodologies. Some of the results have been replicated (e.g. lack of agreement between subjective and statistical techniques) while others have not (e.g. performance of the MHCHISQ complemented by a significance test). Keeping this fact in mind, the results of the various comparative studies reviewed above are

summarized in the next section.

Summary of Findings

Wright (1986) compared the standardization indices and the M-H delta. He manipulated the sample sizes and the number of ability intervals in order to evaluate their effect on the indices. The STD-delta and the M-H delta were very highly correlated thus seeming to detect the same psychometric properties of the items. They correlated highly with each other whether samples were large (>1,000) or small. Wright concluded that because the correlations were not affected by different sample sizes this variable had no effect on them. This conclusion is questionable since the correlations would have still been very high if the sample size affected both indices in a similar manner. The size of the ability intervals does seem to affect the magnitude of the indices since their values were consistently lower when fewer intervals were used.

Five other studies using the M-H procedure were reported above. The MHCHISQ was compared in three of these studies while the M-H delta was used in two. The MHCHISQ compared favorably to the UA when the dif was uniform (Hambleton & Rogers, 1988) but did not perform very well in the Phillips and Mehrens (1988) study. These authors had extremely small samples (87 and 69) in their real data study and as such this part of their work cannot be considered a very good example of the performance of an index. In their simulated study they used 500 subjects per group and the MHCHISQ identified too many items as showing dif. This could be due to the fact that all chi-squares are inflated by sample size. The last study comparing the MHCHISQ to other indices used the

significance test available to evaluate the presence of non-uniform dif (Welch et al, 1987). In this last study, the MHCHISQ complemented by a significance test showed substantial rank order correlations with all the other indices used (i.e. the TID, TID-dis, MCI) except with the SOS3 which did not correlate highly with any of the indices.

The M-H delta was used in two studies: Skaggs and Lissitz (1988) and Dizinno and Arrasmith (1988). The performance of the index in these studies is difficult to evaluate since in the first study, the authors used the upper 20% of the index distribution as an arbitrary cutoff for the indices and in the second study the results of the dif identified by the index were compared to the classification by reviewers. No index performed well in the Skaggs and Lissitz study which could be due to the cutoff used. Little relationship was found between the M-H delta and item reviews (Dizinno & Arrasmith, 1988) which is generally the case in such studies.

As can be seen from these findings, although the M-H procedure is often used today to detect dif, not much is known about its performance. Moreover, little attention has been paid to date to the variables (other than dif) which could influence the MHCHISQ or the M-H delta.

As was mentioned above, the discrimination parameter ("a") influences the M-H delta. The extent of this influence has not been studied. Also, any interaction between the "a" and the difficulty parameter ("b") has also yet to be reported. Furthermore, the effect of sample size which clearly influences the item parameter estimations has not been shown. Other variables may also have some effect on the

M-H (e.g. test length which influences the ability estimations; the pseudo-guessing parameter, "c"). Due to practical considerations, only the "a", "b", and sample size effects will be studied here. The purpose of this study is explained in the next section.

Purpose of this Study

As was mentioned above, the M-H delta is usually preferred over the MHCHISQ for two reasons. First, the MHCHISQ is inflated by sample size and as such should not be used with large samples. Second, The MHCHISQ does not provide a measure of the magnitude of dif as does the M-H delta. For these reasons, only the M-H delta will be studied here.

The purpose of this study is to evaluate the performance of the M-H delta under controlled conditions of the null hypothesis. When the null hypothesis is true, except for random error, the index should not flag any items regardless of other characteristics of the data. Different conditions of the null hypothesis will be simulated in order to evaluate their effects, if any, on the distribution of the M-H delta.

How will the M-H delta perform with different sample sizes? Since using large samples increases the accuracy of any estimation, the M-H delta is expected, with larger samples, to be closer to the value of zero corresponding to the null hypothesis, regardless of any other characteristic(s) of the data. With smaller samples, the accuracy of the index would presumably decrease. This loss of precision should be identified if the index is to be used with small samples.

How will the index perform when the common "a" is equal to one and when it is not? According to equation 5, varying the value of the "a"

parameter should influence the M-H alpha when a difference in "b's" exists. For example, with a "b" difference equal to +.10, the M-H alpha is equal to 1.083 when the "a" equal .8 (M-H delta= -.008), 1.105 with an "a" equal to 1.0 (M-H delta= -.102), and 1.174 when the "a" equals 1.6 (M-H delta= -.164). According to equation 5, the M-H alpha would not be influenced by differing values of the "a" parameter when no difference in "b's" exists in the populations. Random error will nevertheless show differences in "b's" in the samples. To what extent will changing the value of "a" influence the M-H alpha, hence the M-H delta, under the null hypothesis?

Furthermore, fewer examinees for whom the item is appropriate are found at the extremes of the "b" distribution than at its center. The standard errors of the "b" estimates would then be larger for extreme values of "b". This could affect the M-H delta (which should also have larger standard errors). Therefore, does having "b" values at the center or at either extremes of the distribution differentially affect the M-H delta? Does some interaction between sample size, "a", and/or "b" significantly affect the distribution of the M-H delta?

Finally, the cutoff value typically used to flag dif items with the M-H delta is "one" (Wright,1986) (i.e. one quarter of a standard deviation). A cutoff is the index value below which items are not considered to exhibit dif but at and above which the items are flagged as showing dif. Is a cutoff value of one or the value of two (one half of a standard deviation) proposed by P.W. Holland (personal communication, April, 1987) comparable to the cutoffs obtained when no dif is found in the data? Is the usefulness of the cutoffs influenced

by the conditions mentioned above (i.e. sample size and value of "a") or some interaction thereof? How many, if any, false positive identifications of dif items would be made if these cutoffs were used when no dif is induced in the data?

In summary, the following research questions are asked in this study:

1. Is the distribution of the M-H delta affected by sample size, value of "a", and/or value of "b" ?
2. Are the cutoff values affected by the conditions (i.e. sample size, value of "a") or by some combination of the conditions used in this study?
3. How do the cutoffs obtained in this research compare to the values of one proposed by Wright (1986) and that of two suggested by P.W. Holland (personal communication, April, 1987)? How many, if any, false positive identifications would be made in these data if cutoffs of one and two were used?

The present study is designed to answer these questions and its methodology is described in the next chapter.

CHAPTER II

METHODOLOGY

The primary concern in this study was to evaluate the effect, if any, of sample size, item discrimination, and item difficulty on the distribution of the M-H delta. The methodology used in this study is described in the present chapter. This chapter is divided into seven sections which include the following: the data collection approach, the assumptions made in this study, the characteristics of the items, the characteristics of the samples, the simulation model and program, the cutoffs, and the data analysis procedures used in this study.

Data Collection Approach

In order to gather the data necessary to study the problem of interest, a Monte Carlo study was carried out. A simulation study was preferred over using real data because it allows the researcher to simulate known conditions for the study. It provides the researcher the opportunity to choose and manipulate sample size and item parameters. A Monte Carlo study also has the added advantage of allowing replications to be made, thus evaluating the stability of the results.

The general procedure is as follows: the item parameters are fixed to predetermined values; two samples are randomly generated with abilities from a normal (0,1) distribution; response strings are simulated; the M-H index is computed; 100 replications are undergone for each combination of the conditions described in detail below; the M-H index is then analyzed in order to determine the effect, if any, of the different conditions.

Assumptions Made in this Study

Several assumptions were made in this study. First, unidimensionality is assumed. In order to use each examinee's total test score as a measure of ability, it must first be assumed that the total test measures only one ability. If more than one trait were being measured by the test, the total score would not be interpretable. The second assumption concerns the number of items necessary to adequately measure the ability of interest. It was assumed here that 40 items would be sufficient. Third, even if an underlying trait is usually considered to be continuous, dichotomously scored (zero-one coding) items can adequately measure this trait without any significant loss of information. Fourth, it was assumed that all examinees answered all the items, i.e. speed was unimportant. Fifth, no guessing took place. Finally the sixth assumption concerns the independence of responses between the items; that is, the performance on one item does not affect the response to another item.

Characteristics of the Items

A 40 item test was simulated. This test length was chosen because it represents a relatively moderate size test as is commonly found in "real" data. Two item parameters were fixed to predetermined values: the item difficulty ("b") and the item discrimination ("a").

The 40 items were specified for all experimental conditions (described below) with the item difficulties spread across a range from -2.0 to +2.0 with increments of .10. This specific range corresponds to the spread of "b" values most often found in the literature (Hattie, 1984). In order to generate a test with 40 instead of 41 items, the

"b" value of zero was not used.

To study the effect of the discrimination parameter ("a") on the M-H delta, three values of "a" were studied. The first value of "a" chosen for investigation was the value of one implicit in equation 4 (see chapter I). This value was selected in order to study the distribution of the M-H delta when the double assumption of a common "a" equal to one is met. The second and third values of "a" studied here were .7 and 1.3. They were chosen because they represent values at equal intervals from "one" and are often found in real life studies. In each case the value of "a" was common to all items for both "majority" and "minority" groups.

Characteristics of the Samples

In order to study the effect of sample size on the distribution of the M-H delta, three different conditions of this variable were simulated. Since the size of the minority group in "real" data studies is usually smaller than the size of the majority sample, it was decided to use generated "minority" samples that were one third the size of the "majority" group. A similar ratio is found in a number of dif studies using "real" data (e.g. Hambleton & Rogers,1988; Seong & Subkoviak,1987; Shepard, Camilli, & Williams, 1985).

The first sample size chosen here represents a relatively small group: 450 subjects in the "majority" group and 150 in the "minority" sample. Samples smaller than 150 are but rarely used in dif studies. Only one example of smaller sample sizes was found in the literature in a real data study done by Phillips and Mehrens (1988) who used groups of 69 and 87 subjects.

The second and third sample sizes chosen for study were, respectively, two and three times larger than the smallest samples. Doubling and tripling the sample sizes may help evaluate any linear relationship that might exist between sample size and the distribution of the M-H delta. Thus the three sample size conditions generated in this study were as follows: 450/150, 900/300, 1350/450.

Simulation Model and Program

The data were generated by computer using the DATAGEN (Carlson,1983) program. This program can be used in generating data from a one, two, or three parameter model. The user can set specific values for the discrimination ("a"), and the pseudo-guessing ("c") parameters for each of the generated items. The DATAGEN program was modified by L. Fleming (Computer Services, University of Ottawa) in order to set the item difficulties ("b") at a specific value for each item and hold these values constant throughout the replications. The values of "a" and "b" were specified as mentioned above. The DATAGEN program allows the abilities (thetas) to be randomly generated from a normal (0,1) distribution.

The "c" parameter was set to equal zero for all items. This was done for various reasons. The "c" parameter is difficult to estimate even with large samples (Lord, 1980). Furthermore, although this parameter could have an effect on the index studied here, its value is difficult to set with any degree of confidence. For example, with four option multiple choice items, one could set the pseudo-guessing parameter at a value of .25 (i.e. one divided by four options). The choice of this value would nevertheless be difficult to defend. Indeed, a "c" value

of .25 would only reflect a situation where the item was so difficult that all options were equally appealing (or not appealing). In most real life settings, it is felt that examinees who do not know the correct response are attracted to certain of the incorrect options. Thus, guessing occurs less often than one divided by the number of options. For these reasons, and because the interest here was the performance of the M-H delta under the one-parameter logistic model (cf. equations 4 and 5) with and without the assumption of a common "a" equal to one, the "c" parameter was not studied.

The abilities for all simulated subjects were randomly generated from a normal (0,1) distribution. Thus, except for random error, no difference in ability should exist between the groups. Both "majority" and "minority" groups at each level of sample size and value of "a" were generated together starting with a random number. For example, the 600 abilities were generated starting with a random number; the first 450 "subjects" formed the "majority" group, and the last 150 were considered the "minority" group. There were 100 replications of each combination of conditions always using a new random number to start the generation process. The number (100) of replications was chosen in order to adequately represent the distribution of the index within combinations of conditions and according to what was considered feasible.

Once the abilities were simulated, a random number between zero and one was generated for each combination of item and examinee. If this number was equal to or less than the probability of a correct response, the item was scored as correct; otherwise, it was scored as incorrect.

The Cutoffs

When the baseline comparison method, as opposed to a pre-specified value, is used to obtain a cutoff the largest value occurring by chance is usually chosen (e.g. Hambleton & Rogers, 1988; Skaggs & Lissitz, 1988). This means that when samples are equivalent and the index is calculated for each item of a test, the largest value is considered to occur by chance and is used as the cutoff. In order to calculate the cutoffs that would be appropriate for the data generated in this study, the values of the M-H delta at both the $P_{2.5}$ and $P_{97.5}$ of the distribution of the 40 values observed in each replication were computed. These percentiles represent the critical values for a two-tail distribution at a .05 false positive rate. A two-tail distribution was used here since both positive and negative values of the M-H delta represent dif with only the direction of dif changing. Indeed, positive values of the M-H delta correspond to items that the minority group found easier on average, and negative values represent items that the majority group found easier on average. In order to evaluate the value of the index at the .10 false positive rate (and probably with a smaller false negative rate) P_5 and P_{95} were also computed as cutoffs in this study.

These percentiles were compared to the cutoffs suggested in the literature. However, since the program that was used here does not compute M-H delta indices but rather M-H Z indices with a mean of zero and standard deviation of one, the suggested cutoffs were adjusted. That is, a M-H delta of one corresponds to one quarter of a standard deviation on the delta scale, which would be comparable to a M-H Z of

.25; the M-H delta value of two, or one half of a standard deviation on the delta scale, corresponds to a M-H Z value of .50. The false positive rates for .25 and .50 were calculated and compared to cutoffs for the false positive rates set for this study.

Data Analysis

The procedure used to simulate the data in this study consisted of eight steps as follows:

STEP ONE

The item difficulty and discrimination indices were set to specified values. The value of the discrimination depended on the level of that independent variable. The sample size was also determined according to the design of the study. The thetas were then generated followed by each subject's response string.

STEP TWO

Since each generation produced both the "majority" and the "minority" groups, the generated sample was then divided into the prespecified sizes. For example, the program generated 600 thetas for the smallest sample size condition. Dividing this sample produced the required 450/150 split.

STEP THREE

A standardized (0,1) version of the M-H delta (M-H Z) was computed using a program developed by Ackerman (1987).

STEP FOUR

The mean, standard deviation, minimum value, and maximum value of the M-H Z distribution were computed.

STEP FIVE

The chosen cutoffs were computed, that is: $P_{2.5}$, P_{5}, P_{95}, and $P_{97.5}$.

STEP SIX

The number of false positive identifications were computed; that is, the number of items with an index value above the suggested cutoffs of one (M-H Z = $|.25|$) and two (M-H Z = $|.50|$).

STEP SEVEN

One hundred replications were carried out of step one to step six. Each replication started with a different random number.

STEP EIGHT

The data generation process was repeated for each of the nine combinations of sample size and value of "a" conditions.

A 3x3 (sample size by value of "a") analysis of variance (ANOVA), using the SPSSx program, was done on the standard deviation of the M-H delta across the 40 items. The independent variables of sample size and level of "a" should not affect the mean of the M-H Z since its expected value is zero under the null hypothesis. For this reason the ANOVA was not carried out on the mean of the index. On the other hand, if the variables affect the magnitude of the positive and negative values of the M-H Z, this will be reflected in the standard deviation of the index and should be detected by an ANOVA.

Although the standard deviations do not conform to the assumption of a normal distribution underlying an ANOVA, this analysis was shown to be robust against extreme departures from the normal distribution, especially when equal sample sizes are found in each cell as was the case in this study (Godard & Linquist, 1940; Linquist, 1956).

If significance was found, post hoc tests (Scheffe) were carried out. For both the ANOVA and the Scheffe post hoc comparisons a level of significance of .05 was chosen.

A repeated measures ANOVA was carried out on the standard deviation of the M-H Z using the SPSSx program. Since every subject in this study answered all of the 40 items, and each item corresponds to a different value of "b", item difficulty ("b") was a repeated variable. The analysis was undergone on the standard deviation of the index and not on its mean for reasons already mentioned above for the ANOVA. The interest was in studying the effect on the standard deviation of the index of having "b" values close to the center of the distribution as opposed to the extremes, the 40 items were divided into three groups according to their "b" values: "b" values from -2.0 to -0.8 formed the first group; "b" values from -0.7 to +0.6 were included in the second group; and "b" values from +0.7 to +2.0 formed the third group. The level of significance for the repeated measures ANOVA was set at .05.

If a significant effect was found, post hoc comparisons (Scheffe) were carried out. The level of significance used for these analyses was .05.

A MANOVA was undergone on the four cutoffs (percentiles) in order to control for type I error. The multivariate F test used to evaluate the significance of the results was the Pillai-Bartlett since it is considered to be very robust against violations of the assumptions of multivariate normality, homogeneity of variance, and linearity underlying a MANOVA analysis (Barcikowski, 1983). For this analysis the level of significance was set at .05. If significant, univariate

F tests, and Scheffe post-hoc comparisons were carried out. To control for type I error, for the univariate F tests and the Scheffe tests, the significance level was .013, that is, .05 divided by four dependent variables. All the analyses in this study were undergone using the SPSSx program.

The number and proportion of false positive identifications made by the .25 and .50 cutoffs were computed for each combination of conditions. These cutoffs were then compared to those identified for the false positive rates from this study.

Results and accompanying discussion are the topic of the next chapter.

CHAPTER III

RESULTS AND DISCUSSION

In this chapter the results of this study are presented and discussed. This chapter is divided in five sections as follows: distribution of the M-H Z, effect of the independent variables on the M-H Z, effect of the independent variables on the cutoffs, analysis of the false positive identifications, and summary of findings.

Distribution of the M-H Z

The mean and standard deviation of the mean of the distribution of the M-H Z over 100 replications were computed and are displayed by conditions in Table 1. These values were computed across the 40 items, that is, across the 40 different values of "b".

Table 1

Mean and Standard Deviation of the 100 Means of the M-H Z

	Mean	Standard Deviation of the Mean
N = 450/150		
a = 0.7	.001	.006
a = 1.0	.000	.011
a = 1.3	-.002	.016
N = 900/300		
a = 0.7	.000	.004
a = 1.0	.001	.006
a = 1.3	.000	.009
N = 1350/450		
a = 0.7	.000	.000
a = 1.0	.001	.000
a = 1.3	.000	.006

As can be seen in Table 1, the mean of the M-H Z was always around the expected value of .000. The standard deviations varied somewhat

with sample size and value of "a". Increasing the sample size decreased the standard deviation of the mean of the distribution of the M-H Z which was expected since larger samples should produce more stable estimates. Increasing the "a" also had the effect of increasing the standard deviation of the index. In other words when the discrimination parameter was higher, the mean of the M-H Z was less stable.

The effect of sample size and value of "a" on the standard deviation of the index was also clear when the standard deviation was computed for each of the 40 items (values of "b") across the 100 replications. Table 2 shows these values where the same pattern was observed; that is, increasing sample size decreases the standard deviation, and increasing the "a" inflates the standard deviation. From Table 2 it can also be observed that items with "b" values at either extreme of the distribution tended to have higher standard deviations, hence a less stable M-H Z, showing more variability over 100 replications than did items with "b" values closer to the center of the distribution.

Table 2

Standard Deviation of the M-H Z for each Value of "b" across 100 Replications.

	N=450/150			N=900/300			N=1350/450		
"b"	"a"			"a"			"a"		
	0.7	1.0	1.3	0.7	1.0	1.3	0.7	1.0	1.3
-2.0	.172	.262	.451	.138	.194	.214	.100	.152	.167
-1.9	.190	.195	.324	.114	.154	.175	.093	.124	.161
-1.8	.207	.249	.319	.117	.147	.171	.094	.123	.131

Table 2 (continued)

	N-450/150			N-900/300			N-1350/450		
"b"	"a"			"a"			"a"		
	0.7	1.0	1.3	0.7	1.0	1.3	0.7	1.0	1.3
-1.7	.198	.237	.276	.133	.150	.172	.106	.121	.140
-1.6	.164	.238	.204	.129	.141	.169	.091	.110	.137
-1.5	.141	.200	.280	.115	.152	.156	.083	.099	.125
-1.4	.156	.202	.250	.115	.124	.165	.089	.114	.140
-1.3	.146	.152	.212	.100	.127	.138	.095	.103	.120
-1.2	.173	.193	.196	.116	.116	.135	.089	.105	.109
-1.1	.138	.173	.183	.097	.116	.135	.091	.074	.094
-1.0	.136	.188	.202	.099	.128	.142	.075	.100	.088
-0.9	.141	.151	.181	.110	.129	.112	.086	.089	.109
-0.8	.133	.167	.159	.094	.104	.125	.085	.083	.116
-0.7	.122	.173	.161	.093	.103	.119	.077	.095	.102
-0.6	.132	.146	.177	.097	.109	.112	.070	.080	.091
-0.5	.144	.157	.159	.088	.102	.119	.070	.092	.098
-0.4	.137	.153	.173	.090	.093	.120	.074	.083	.103
-0.3	.132	.164	.152	.092	.107	.111	.081	.088	.093
-0.2	.135	.142	.149	.098	.108	.099	.066	.073	.094
-0.1	.133	.136	.165	.087	.091	.113	.077	.076	.089
+0.1	.134	.136	.164	.092	.096	.113	.064	.083	.089
+0.2	.137	.151	.163	.089	.098	.112	.067	.078	.086
+0.3	.131	.145	.124	.094	.102	.113	.075	.081	.078
+0.4	.140	.143	.161	.101	.111	.111	.079	.094	.087
+0.5	.129	.133	.172	.087	.113	.109	.088	.082	.087

Table 2 (continued)

	N=450/150			N=900/300			N=1350/450		
"b"	"a"			"a"			"a"		
	0.7	1.0	1.3	0.7	1.0	1.3	0.7	1.0	1.3
+0.6	.157	.170	.184	.095	.114	.105	.076	.093	.086
+0.7	.133	.155	.163	.101	.107	.114	.082	.084	.095
+0.8	.164	.164	.155	.105	.121	.126	.076	.090	.115
+0.9	.136	.156	.198	.113	.124	.136	.083	.090	.108
+1.0	.155	.165	.198	.103	.126	.123	.077	.091	.104
+1.1	.128	.168	.225	.112	.122	.146	.104	.101	.118
+1.2	.164	.158	.214	.109	.119	.124	.087	.097	.130
+1.3	.164	.198	.216	.119	.116	.139	.092	.097	.136
+1.4	.193	.170	.193	.109	.121	.151	.089	.103	.117
+1.5	.167	.210	.262	.108	.157	.137	.102	.115	.127
+1.6	.185	.195	.269	.123	.139	.163	.090	.111	.124
+1.7	.184	.215	.262	.125	.138	.179	.093	.123	.150
+1.8	.212	.216	.270	.134	.159	.177	.100	.120	.154
+1.9	.203	.210	.302	.125	.141	.192	.124	.133	.158
+2.0	.197	.238	.338	.134	.167	.206	.094	.130	.155

The standard deviations were also computed across the 40 items for each of the 100 replications of each combination of conditions. The distribution of this statistic, that is the mean, the standard deviation, the minimum, and the maximum values of the standard deviations are shown in Table 3.

As can be seen in Table 3, the mean standard deviation of the

distribution of the M-H Z over the 100 replications increased when the value of "a" increased but decreased with larger sample sizes, ranging from .087 for a = 0.7 and largest sample size to .222 for a = 1.3 and smallest sample size.

Table 3

Distribution of the Standard Deviation (SD) of the M-H Z

Conditions	Mean of the SD	Standard deviation of the SD	Min	Max
N = 450/150				
a = 0.7	.159	.021	.100	.223
a = 1.0	.183	.022	.143	.253
a = 1.3	.222	.050	.153	.588
N = 900/300				
a = 0.7	.109	.015	.171	.610
a = 1.0	.127	.017	.090	.165
a = 1.3	.143	.016	.108	.186
N = 1350/450				
a = 0.7	.087	.010	.061	.112
a = 1.0	.101	.013	.075	.139
a = 1.3	.118	.014	.089	.165

Effect of the Independent Variables on the M-H Z

To determine the effect of sample size and value of "a" on the standard deviation of the M-H Z across the 40 items, an analysis of variance (ANOVA) was carried out. A three by three ANOVA (i.e. three

sample sizes by three values of "a") was used with the dependent variable being the standard deviation computed across the 40 items. Each combination of conditions thus included 100 standard deviations, one for each replication. The results of the ANOVA are shown in Table 4.

Table 4

Results of the 3x3 ANOVA on the Standard Deviation of the M-H Z across the 40 Items.

Source	SS	df	MS	F	P
Sample size	1.168	2	.584	1122.82	.000*
"a" value	.276	2	.138	265.52	.000*
Sample size by "a" value	.036	4	.009	17.27	.000*
Residual	.464	891	.001		

* = significant at the .05 level

As can be seen in Table 4, the interaction between sample size and value of "a" was significant at the .05 level. Therefore one-way ANOVA's for each of the two independent variables (sample size or "a") when the other variable ("a" or sample size) was held constant, and Scheffe post hoc comparisons were carried out in order to determine which level(s) of the variables differed significantly.

The detailed results of the simple effects testing are shown in appendix A (see Table 16). All pairs of each independent variable's levels were significantly different. In other words, the sample size

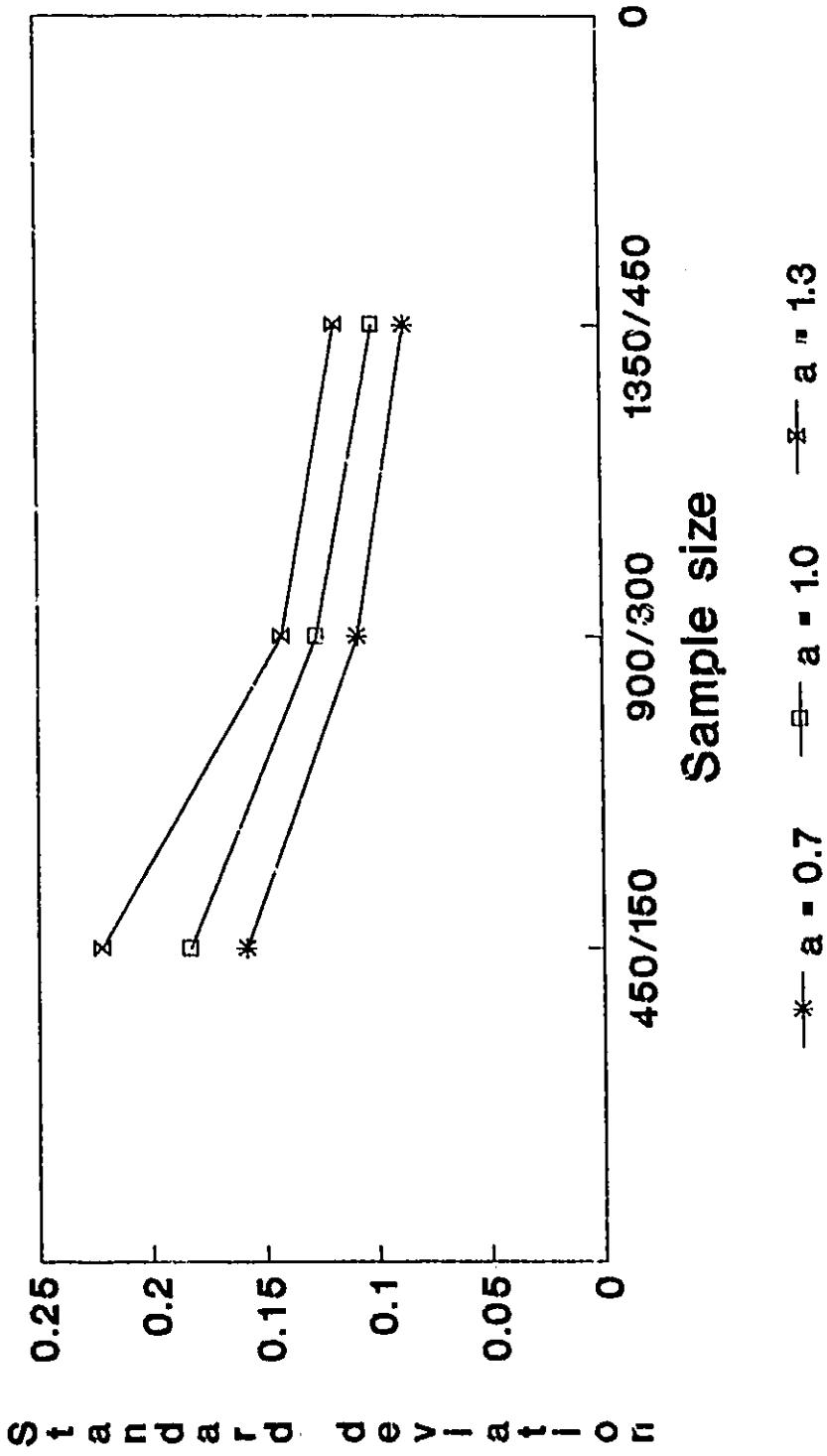
had a significant effect for each pair of values of "a" and value of "a" had a significant effect on the standard deviation of the M-H Z for each pair of sample size.

The two-way interaction was plotted and is shown in Figure 1. As seen from Figure 1, the interaction is not disordinal. As sample size increased, the standard deviation decreased and as the value of "a" increased, so did the standard deviation. The interaction appears to be explained by the more than additive effect of large "a" and small sample size.

One of the research questions to be answered concerned the effect on the distribution of the M-H delta (or the M-H Z) of the location of the difficulty parameter ("b"). The M-H Z was expected to be more stable when the "b" value was located close to the center of the distribution than when the "b's" were at either extreme. In order to answer this question, the 40 values of "b" were not studied separately. Rather, three different groups of "b" values were analyzed.

In the first group were "b" values between -2.0 and -0.8; in the second group, values from -0.7 to +0.6; in the last group were "b" values from +0.7 to +2.0. These three groups represent approximately equal intervals of "b" values (1.2, 1.3, and 1.3, respectively) and number of items (13, 13, and 14, respectively). The effect of these variables (sample size, value of "a", group of "b") should not be significant on the mean of the M-H Z. This was confirmed when the means of the M-H Z were examined. They ranged from -.008 to +.009 for the 27 combinations of conditions. Hence only the standard deviations of the M-H Z were analyzed.

Figure 1: Standard Deviation
Sample Size by "a" Interaction



A repeated measures ANOVA was undergone on the standard deviations of the M-H Z across the 100 replications of each combination of conditions. The assumption of sphericity was not violated since both the Greenhouse-Geisser and the Huynh-Feldt epsilons were over .90. The results of the repeated measures ANOVA are displayed in Table 5. Table 5 shows the sample size by value of "a" had a significant effect of the standard deviation of the M-H Z.

Table 5

Repeated Measures ANOVA on the Standard Deviation of the M-H Z

BETWEEN SUBJECTS

Source	SS	df	MS	F	P
Sample size	3.32	2	1.66	1336.14	.000*
"a"	.09	2	.38	308.77	.000*
Sample size by "a"	.09	4	.02	17.15	.000*
Within cells	1.11	891	.00		

WITHIN SUBJECTS

Source	SS	df	MS	F	P
"b" values	.82	2	.41	353.98	.000*
Sample size by "b" values	.08	4	.02	16.18	.000*
"a" by "b" values	.10	4	.03	21.94	.000*
Sample size by "a" by "b" values	.02	8	.00	2.50	.000*
Within cells	2.07	1782	.00		

* = Significant effect at the .05 level

This analysis was undergone with the means of the standard

deviations of each of the three groups of "b". The effect of the sample size by value of "a" interaction is the same as was seen in the previous analysis done with the standard deviation computed over the 40 items. For this reason no follow up tests were done and the results will not be discussed further.

The three-way interaction between sample size, value of "a", and value of "b" was also shown to be significant (alpha = 0.05). To assess this interaction, a simple repeated measures ANOVA was carried out with sample size and value of "a" held constant. This was followed by Scheffe post hoc comparisons. The results are reported in detail in appendix A (see Table 17). Table 6 shows the means over 100 replications of the standard deviations across each group of "b" values and for each combination of sample size and value of "a".

Table 6

Means over 100 Replications of the Standard Deviation across Groups of "b" and for each Combination of Sample Size and Value of "a" Conditions.

	Groups of "b" values		
	-2.0 to -0.8	-0.7 to +0.6	+0.7 to +2.0
N = 450/150			
a = 0.7	.1596	.1353	.1723
a = 1.0	.2007	.1479	.1904
a = 1.3	.2413	.1631	.2429
N = 900/300			
a = 0.7	.1122	.0926	.1163
a = 1.0	.1355	.1029	.1339
a = 1.3	.1542	.1124	.1556
N = 1350/450			
a = 0.7	.0895	.0747	.0935
a = 1.0	.1083	.0831	.1073
a = 1.3	.1257	.0905	.1331

In all combinations of sample size and value of "a", the standard deviations were significantly larger when the items' difficulties were located at either extremes of the "b" distribution than when the "b's" were in its center but did not differ significantly between the two extremes.

Figures 2 to 4 show the effect of groups of "b" values on the standard deviation when sample size was held constant. These figures show quite clearly the two- way interaction between "a" and "b": the effect of the "a" was more pronounced when the "b" was at either extreme of the distribution. The three-way interaction is explained by the more than additive effect when "a" equalled 1.3, "b" was at either extreme, and sample sizes were small (figure 2).

Effect of the Independent Variables on the Cutoffs

Four cutoffs were computed: $P_{2.5}$, P_5, P_{95}, and $P_{97.5}$. These values correspond to different false positive identification (FPI) rates, that is, $P_{2.5}$ and $P_{97.5}$ represent a FPI of .05 and P_5 and P_{95} correspond to a FPI of .10.

The means and standard deviations over 100 replications of each of these percentiles were computed for each combination of conditions and are shown in Table 7.

The means of the cutoffs are affected by both sample size and value of "a". As sample size increased, the absolute values of the percentiles decreased and as value of "a" increased, the absolute values of the cutoffs increased.

Figure 2:"b" by "a" Interaction
Sample Size = 450/150

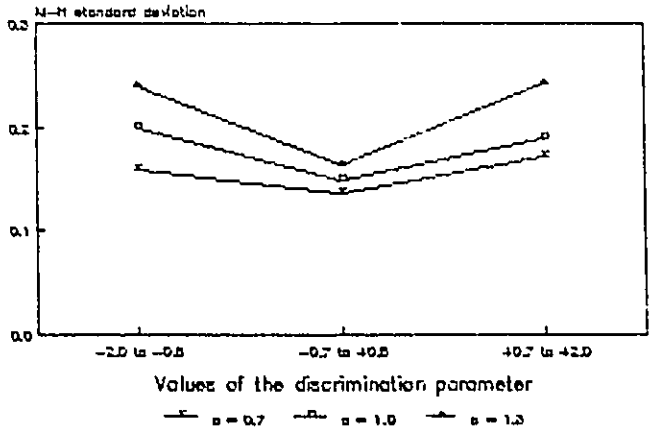


Figure 3:"b" by "a" Interaction
Sample Size = 900/300

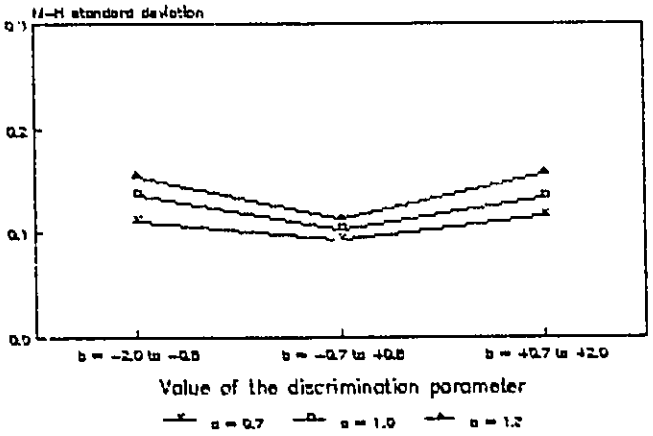


Figure 4:"b" by "a" Interaction
Sample size = 1350/450

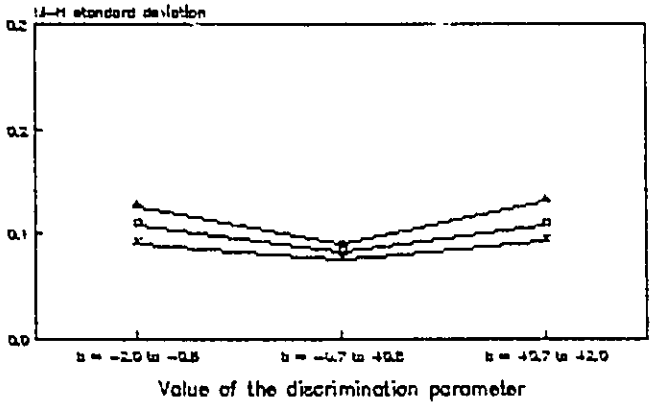


Table 7

Means and Standard Deviations of the Cutoffs at each Combination of Conditions.

	$P_{2.5}$	$P_{97.5}$	P_5	P_{95}
	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)
N = 450/150				
a = 0.7	-.364 (.088)	.346 (.091)	-.284 (.064)	.276 (.048)
a = 1.0	-.413 (.092)	.420 (.111)	-.316 (.059)	.331 (.077)
a = 1.3	-.517 (.160)	.540 (.345)	-.392 (.077)	.380 (.089)
N = 900/300				
a = 0.7	-.240 (.066)	.240 (.052)	-.189 (.041)	.190 (.041)
a = 1.0	-.283 (.069)	.286 (.079)	-.221 (.041)	.222 (.048)
a = 1.3	-.325 (.089)	.327 (.086)	-.252 (.050)	.253 (.053)
N = 1350/450				
a = 0.7	-.200 (.054)	.186 (.039)	-.154 (.033)	.152 (.030)
a = 1.0	-.218 (.052)	.230 (.064)	-.172 (.033)	.185 (.040)
a = 1.3	-.269 (.059)	.265 (.058)	-.210 (.042)	.207 (.042)

The two most extreme values in Table 8 were found for $P_{2.5}$ and $P_{97.5}$ when sample sizes were small (450/150) and "a" was large (1.3). Notice that the standard deviations for these values are the largest. This is due to the fact that both these percentiles had extreme values, -1.151 for $P_{2.5}$ and 3.370 for $P_{97.5}$, that account for the inflated means and standard deviations.

To assess the stability of the cutoffs, four confidence intervals were computed: first these were calculated for the two largest absolute

cutoff values found in Table 8 which were observed when samples were small (450/150) and "a" was equal to 1.3; then, confidence intervals were computed for the two smallest absolute values of Table 8, observed when sample were large (1350/450) and "a" was equal to 0.7.

The largest values were for $P_{2.5}$ (-0.517 with a standard deviation of 0.160) and for $P_{97.5}$ (0.540 with a standard deviation of 0.345). The .95 confidence intervals would then be as follows:

$$-0.830 < P_{2.5} < -0.203$$

$$0.136 < P_{97.5} < 1.216$$

The smallest absolute cutoff values were for P_5 (-.154 with a standard deviation of 0.033) and for P_{95} (.152 with a standard deviation of 0.03). The .95 confidence intervals would be as follows:

$$-0.219 < P_5 < -0.089$$

$$-0.093 < P_{95} < 0.211$$

These intervals are fairly large and indicate that these cutoffs are somewhat unstable. In view of this instability the danger of using one value as a cutoff to identify dif by the baseline comparison method must be stressed. These unstable cutoffs may inappropriately classify items as exhibiting dif or not.

In order to examine the cutoffs without the effect of the extreme values, their medians were computed along with the semi-interquartile range. The results are shown in Table 8. As expected, the medians are somewhat lower than the means shown in Table 7. Nevertheless, the pattern of sample size and "a" value effect remained unchanged. Notice that, for all sample sizes, the semi-interquartile ranges were similar when "a" equalled 0.7 or 1.0 but were higher when "a" equalled 1.3.

Table 8

Median (and semi-interquartile range) for each Cutoff at each Combination of Conditions.

	$P_{2.5}$	$P_{97.5}$	P_5	P_{95}
N = 300/150				
a = 0.7	-.353 (.059)	.324 (.049)	-.278 (.043)	.269 (.032)
a = 1.0	-.408 (.065)	.398 (.068)	-.302 (.045)	.322 (.040)
a = 1.3	-.479 (.089)	.460 (.087)	-.396 (.058)	.371 (.059)
N = 900/300				
a = 0.7	-.229 (.045)	.234 (.036)	-.179 (.024)	.187 (.027)
a = 1.0	-.279 (.043)	.269 (.045)	-.218 (.029)	.218 (.024)
a = 1.3	-.308 (.053)	.318 (.054)	-.246 (.030)	.239 (.035)
N = 1350/450				
a = 0.7	-.189 (.030)	.179 (.025)	-.148 (.020)	.149 (.016)
a = 1.0	-.209 (.029)	.219 (.033)	-.169 (.024)	.187 (.023)
a = 1.3	-.269 (.040)	.258 (.031)	-.209 (.029)	.200 (.029)

A MANOVA was carried out on the four cutoffs. The results of this analysis are shown in Table 9. As can be seen from Table 9, sample size, discrimination, and the interaction of sample size and discrimination are all significant at the .05 level. Since the multivariate test was significant, univariate F tests were carried out in order to determine whether the interaction occurred for each of the four cutoffs.

Table 9

Multivariate Analysis of the Cutoffs over 100 Replications

Source	Pillai-Bartlett	df	Error df	P
Sample size	.72635	8	1778	.000*
"a"	.39844	8	1778	.000*
Sample size by "a"	.07651	16	3564	.000*

* = significant at the .05 level

Table 10

Univariate F tests for the Sample Size by Value of "a" Interaction Effect on the Cutoffs

Variable	SS	MS	Error MS	F	P
$P_{2.5}$	.22448	.05612	.00757	7.42	.000*
P_{5}	.09663	.02416	.00260	9.30	.000*
P_{95}	.07267	.01817	.01385	6.01	.000*
$P_{97.5}$	.44582	.11146	.01385	6.07	.000*

* = significant at .013, F(4,891)

Table 10 shows the results of the univariate F test for the sample size by discrimination value interaction effect. Since the level of significance here was .013 (i.e. .05 divided by four), all cutoffs can be seen to show significant results.

Thus it can be concluded that each of the cutoffs (percentiles) is significantly affected by the interaction between sample size and value of "a". Thus for each cutoff, one-way ANOVA's were carried out for

each of the two independent variables when the other variable was held constant. These were followed by Scheffe post hoc comparisons.

The first simple effect testing and post hoc tests that were done concern the sample size effect with the discrimination value held constant at 0.7, 1.0, and 1.3. If the simple effect for one percentile was significant at the .013 level (i.e. .05 divided by four percentiles) then differences between the means at the three levels of sample size were computed and Scheffe tests (at the .013 level of significance) carried out to determine between which level(s) of sample size the significant differences occurred. The results of the one-way ANOVA's and the post hoc tests are shown in tables 18 through 23 of appendix A. Table 11 summarizes the results of the post hoc comparisons carried out on the four percentiles for the sample size effect when the values of "a" are held constant.

For all percentiles, that is, for $P_{2.5}$, P_{5}, P_{95}, and for $P_{97.5}$ sample size had a significant effect (at alpha = .013) when the discrimination parameter was held at 0.7. Scheffe post hoc comparisons revealed that all groups differed significantly from one another; that is, for all percentiles, the difference between having a smaller sample (N = 450/150) and having any of the two larger samples is significant, as is the difference between having a relatively moderate size sample (N = 900/300) and a larger group (N = 1350/450). It can be concluded that sample size, at all the levels studied here, had a significant effect on all the cutoffs when the discrimination parameter was equal to 0.7.

With the discrimination parameter equal to one, sample size had a significant effect on the four percentiles and each cutoff differed

Table 11

Summary of the Post Hoc Tests on the Four Percentiles for the Sample Size Effect when the Values of "a" are held Constant.

	SAMPLE SIZE EFFECT *			
	$P_{2.5}$	$P_{97.5}$	P_{5}	P_{95}
"a" = 0.7	1,2	1,2	1,2	1,2
	1,3	1,3	1,3	1,3
	2,3	2,3	2,3	2,3
"a" = 1.0	1,2	1,2	1,2	1,2
	1,3	1,3	1,3	1,3
	2,3	2,3	2,3	2,3
"a" = 1.3	1,2	1,2	1,2	1,2
	1,3	1,3	1,3	1,3
	2,3	2,3ns	2,3	2,3

* where 1 : N = 450/150
2 : N = 900/300
3 : N = 1350/450

ns = groups are not significantly different

significantly for all pairs of sample sizes.

Sample size had a significant effect on all the percentiles studied when the discrimination was equal to 1.3. For $P_{2.5}$, P_{5}, and P_{95} all three levels of sample size differed significantly with one another. For $P_{97.5}$, however, a significant difference existed only between the smallest sample size (N=450/150) and the two others. This means that all three levels of sample size had an effect on the first three percentiles, but that with the last cutoff, using moderate or larger samples did not make a difference while using smaller samples did.

Table 12 summarizes the results of the post hoc tests undergone on the discrimination effect when the sample

size was held constant.

Table 12

Summary of the Post Hoc Tests on the Four Percentiles for the Discrimination Effect when the Sample Size is held Constant

	DISCRIMINATION EFFECT *			
	$P_{2.5}$	$P_{97.5}$	P_5	P_{95}
N - 450/150	1,2	1,2 ns	1,2	1,2
	1,3	1,3	1,3	1,3
	2,3	2,3	2,3	2,3
N - 900/300	1,2	1,2	1,2	1,2
	1,3	1,3	1,3	1,3
	2,3	2,3	2,3	2,3
N - 1350/450	1,2 ns	1,2	1,2	1,2
	1,3	1,3	1,3	1,3
	2,3	2,3	2,3	2,3

* where 1 : "a" - 0.7
2 : "a" - 1.0
3 : "a" - 1.3

ns - groups are not significantly different

When sample size was small, the discrimination parameter had a significant effect on all the percentiles. For $P_{2.5}$, P_5, and P_{95} there were significant differences between all levels of the discrimination parameter. For $P_{97.5}$, there was no significant difference between having the "a" equal to 0.7 and 1.0, but a significant difference did occur between each of these two values of "a" and a value of 1.3. In other words, the discrimination parameter does affect the percentiles. All the values of "a" studied here affected all of the percentiles, except for the $P_{97.5}$ which did not differ significantly when "a" went from 0.7

to 1.0.

When the samples were equal to 900/300, the discrimination effect was significant for all the percentiles studied. Furthermore, each percentile differed significantly when the discrimination went from 0.7 to 1.0 or to 1.3, and from 1.0 to 1.3.

When samples were large (1350/450), the discrimination parameter value had a significant effect on the four percentiles studied. For P_5, P_{95}, and $P_{97.5}$, all pairs of the discrimination variable were significantly different from each other. For $P_{2.5}$, there were no differences between "a" levels of 0.7 and 1.0.

It can be said that, basically, the sample size by value of "a" interaction effect is explained as follows: with small sample sizes (450/150) the absolute values of the cutoffs are inflated by larger values of "a". As with the other interactions studied here between sample size and value of "a", a more than additive effect occurs with combinations of small sample size and large "a".

Analysis of the False Positive Identifications

Two cutoffs have been suggested to identify dif items. One corresponds to one quarter of a standard deviation of a standardized "p", i.e. a value of .25 for the M-H Z. The second cutoff represents one half of a standard deviation, that is, .50 for the index of interest here.

One of the research questions concerned the number of false positive identifications that would be made if these cutoffs were used with the data generated in this study. Since no dif was induced in these data, any item with an index value at or above the cutoff would be wrongly

flagged as exhibiting dif. Table 13 shows the number of such false positive identifications (along with their respective proportions) that would be made if a cutoff of .25 were used and with a cutoff of .50.

Table 13

Number of False Positive Identifications (and Respective Proportions) by Sample Size and "a" Value.

(number of items times replications equals 4,000)

	CUTOFFS	
	|.25|	|.50|
N= 450/150		
a =0.7	388 (.097)	13 (.003)
a =1.0	573 (.143)	44 (.011)
a =1.3	616 (.154)	118 (.030)
Subtotal	1577 (.131)	175 (.01)
N= 900/300		
a =0.7	115 (.029)	1 (.0003)
a =1.0	201 (.050)	4 (.001)
a =1.3	302 (.076)	11 (.003)
Subtotal	618 (.052)	16 (.001)
N= 1350/450		
a =0.7	28 (.007)	0
a =1.0	68 (.017)	0
a =1.3	168 (.042)	0
Subtotal	264 (.022)	0

As Table 13 shows, for .25, more false positive identifications (FPI) would have been likely to occur with a small sample size (1577) and as the sample size increases, the number of FPI decreases (to 618 and then, 264). When sample size was held constant, more false

positive identifications were made when the "a" value was large. This effect was also noticeable when the .50 cutoff was used.

The proportions shown in table 13 can be used to evaluate the appropriateness of the cutoffs for a .05 false positive rate. In this table, the sign of the index was not taken into account, that is, an item was flagged as dif by the .25 cutoff, for example, if it had an absolute M-H Z value at or above .25.

With the .50 cutoff, no condition showed proportions of FPI higher than .05. That is to say, with this cutoff too few items would be flagged as dif which may affect the number of false negative identifications that would be made if there was dif in the data. When samples were small (450/150), on the other hand, the .25 cutoff seemed too lenient, that is, flagging too many items, regardless of the "a" value.

With moderate samples (N=900/300), the .25 cutoff flagged approximately 5% of the items which is what is expected of a false positive rate of .05.

With larger samples, the overall proportion of FPI with the .25 cutoff was .022. In this last case, when the "a" value was equal to 1.3, the proportion of FPI's was close to the .05 level (.042). But the proportions of FPI's with "a" equal to 0.7 and 1.0 seem likely to be too low (.007 and .017, respectively) since low proportions of FPI may be accompanied with higher proportions of false negative identifications.

Since the repeated measures ANOVA reported above (Table 5) found that the values of "b" affected the standard deviation of the M-H Z,

the number of false positive identifications were computed for each group of "b" values. Only the .25 cutoff is shown in Table 14 since the .50 cutoff was believed to be too stringent for these data, and would probably allow too many false negative identifications.

Table 14

False Positive Identifications (and Proportions) by Conditions and by Grouped Values of "b".

	Values of "b"		
"b" =	-2.0 to -.8	-.7 to +.6	.7 to 2.0
N = 450/150			
a = 0.7	158 (.12)	91 (.07)	139 (.10)
a = 1.0	222 (.17)	139 (.11)	212 (.15)
a = 1.3	246 (.19)	143 (.11)	227 (.16)
subtotal	626 (.16)	373 (.10)	578 (.14)
N = 900/300			
a = 0.7	44 (.03)	10 (.01)	61 (.04)
a = 1.0	97 (.07)	22 (.02)	82 (.06)
a = 1.3	122 (.09)	48 (.04)	132 (.09)
subtotal	263 (.07)	80 (.02)	275 (.07)
N = 1350/450			
a = 0.7	12 (.01)	3 (.002)	13 (.009)
a = 1.0	34 (.03)	0	34 (.02)
a = 1.3	76 (.06)	9 (.007)	83 (.06)
subtotal	122 (.03)	12 (.003)	130 (.06)

Table 14 provides a clear indication of the effect on the M-H Z of the location of the "b" value in the distribution. When the difficulty was near the center of the distribution, fewer false positive

identifications were made with all sample sizes and across all values of the discrimination parameter.

To assess the cutoffs that would be appropriate for these data, the absolute value of the means and of the medians of the four percentiles studied here were used. Both the means and the medians were used in order to evaluate the values observed, respectively, with and without the influence of extreme values. P_5 and P_{95} represent a .10 false positive rate while the values of $P_{2.5}$ and $P_{97.5}$ correspond to a .05 false positive rate. Table 15 lists these values.

Table 15

Absolute Values of the Mean (and the Median) of the Percentiles Corresponding to a .10 and to a .05 False Positive Rate.

	False Positive Rates	
	.10	.05
N=450/150	Mean (Median)	Mean (Median)
a = 0.7	.280 (.274)	.350 (.339)
a = 1.0	.323 (.312)	.415 (.403)
a = 1.3	.386 (.384)	.528 (.470)
N=900/300		
a = 0.7	.189 (.183)	.240 (.232)
a = 1.0	.221 (.218)	.285 (.274)
a = 1.3	.252 (.243)	.326 (.313)
N=1350/450		
a = 0.7	.153 (.149)	.193 (.184)
a = 1.0	.178 (.178)	.224 (.214)
a = 1.3	.208 (.205)	.267 (.264)

The values shown in Table 15, especially the values corresponding to the .05 level, represent the cutoffs that would be appropriate to use with the data in this study.

The summary of the findings along with the limitations of this study and suggestions for further research are found in the next chapter.

CHAPTER IV

CONCLUSION

This concluding chapter is divided into three sections: summary of findings, limitations of this study, and suggestions for further research.

Summary of Findings

The two-way interaction between sample size and value of "a" was significant for the standard deviation of the M-H Z distribution. There was nothing disordinal about this interaction. Increasing sample size always significantly decreased the standard deviation, regardless of the value of "a", while increasing the value of "a" always significantly inflated the standard deviation regardless of sample size. However, when samples were small (450/150), having large "a's" inflated the standard deviation noticeably more and the interaction was explained by the exaggerated effect of this combination of conditions.

Thus it can be concluded that with the data used here the variability of the M-H Z (or M-H delta) was significantly larger when the "a" increased and it was significantly lower as sample size increased.

The repeated measures analysis replicated the findings described above with the added results that when the "b" values were located near the center of the distribution, the standard deviation of the M-H Z was significantly lower than when the "b's" were found at either extremes. In the repeated measures analysis, the three-way interaction was significant: the two-way interaction between "a" and "b" was explained by the smaller effect of "a" (especially large "a's") when "b" was

located at the center of the distribution; the three-way interaction was due to the exaggerated effect of the two-way interaction when samples were small.

The effect of sample size by "a" interaction was significant in the MANOVA carried out on the four percentiles studied. That is, larger sample sizes always deflated the percentiles while large "a's" inflated them, but this effect was exaggerated when samples were small and "a's" were large.

The lack of stability of the cutoffs over replications puts in question the advisability of using the baseline comparison method to determine the cutoff used in a study. Since the cutoffs varied significantly from one combination of conditions to the next, choosing the highest value that occurs in any one replication could lead to dif classification errors. The cutoffs also differed greatly within combination of conditions, so that even if one knew the effect of sample size and discrimination, the cutoffs would still differ noticeably from one replication to another.

The number of false positive identifications that would be made when the cutoffs of .25 and .50 were used seemed to be influenced by both sample size, value of "a", and value of "b". Increasing the sample size decreased the number of false positive identifications. Increasing the "a" parameter inflated the number of false positive identifications. The absolute value of the "b" parameter also seemed to influence the number of false positive identifications. That is, when the value of "b" was near the center of the distribution, the number of false positive identifications was considerably lower than

when the "b" was located at either extreme.

In conclusion, the results of the present study put in question the usefulness of the M-H delta without taking the sample size, value of "a", and value of "b" into account. This index has been suggested for use with small or large samples, which would have been a clear advantage of this index over the item response theory indices which need large sample sizes. The lack of stability of the M-H delta across sample size was shown in this study. Its standard deviation is influenced by all the conditions studied here, that is, sample size, value of "a", and value of "b". In addition, the large range of the confidence intervals computed in this study put in question the use of only one replication to choose a cutoff.

In order for the M-H delta to be an adequate dif detection index, it should not be influenced by variables other than dif. The fact that sample size, value of "a", and value of "b" all have significant effects on the index greatly diminishes the confidence that can be put on the M-H delta. Since tests developers do not have as much knowledge about their data as was the case in this study where all the independent variables were controlled, and since appropriate cutoffs appear to depend on sample size, value of "a", value of "b", and replication identifying items as showing dif by one given value of the M-H delta would seem to be inadvisable. The practitioner's problem would lie in not knowing which of the variables inflated the index. The M-H delta, as it is now commonly used, does not appear to be useful in identifying dif and only dif.

Limitations of this Study

The greatest advantage of this study, that is, the control of the independent variables, is also its greatest limitation. Although an effort was made to generate data that would simulate real life situations, no assurance can be made that the data described in this study compare to "real" data where many variables other than the ones studied here, such as test length, probably influence the results. Longer tests would presumably give more accurate ability estimates which could affect the index.

Furthermore, the restrictions applied on the data may have influenced the results. Choosing sample sizes, values of the item discrimination, and values of the item difficulty that would differ from those used in this study may alter the results although the findings of this study may allow some prediction to take place. No guessing was allowed to occur here though this variable could very well influence the results. Another variable that could affect the results may be different rates of completion, i.e. when not all examinees answer all the items in the test.

All the restrictions put on these data could have affected the results to some extent. The findings of this study suggest that further research may be needed.

Suggestions for Further Research

The suggestions for future research given here stem directly from the limitations of this study. Since the results of a study can be generalized only to similar data sets, the results found here should be replicated with different data before being accepted.

For instance, a similar study could be undergone with "real" data. Different sample sizes could be used. The item difficulty and item discrimination could then be computed (if the data bank was large enough) and used to evaluate their effect on the M-H delta. The pseudo-guessing parameter could also be estimated in order to study its effect on the index. Different test lengths could be used.

Further simulation studies could also be undergone where these variables would be more easily manipulated. Further studies could be conducted to evaluate the proportion of false negative identifications that would be made when dif was induced in the data. The effect of sample size, value of "a", and value of "b" on the proportion of false negative identifications is an important issue to consider if the M-H delta is to be used to identify dif.

If the results of these studies replicated the ones found here, then the M-H delta, as it is commonly used now, would have to be considered too unstable for use in identifying dif items. The cutoff used to identify dif items would have to be chosen according to sample size, value of "a", and value of "b". Since the M-H procedure is increasingly popular today, the results of the present study are disturbing and should be replicated in order to evaluate their generalizability. Practitioners in education should be aware that variables other than dif affect the M-H delta and that the cutoffs should be chosen accordingly.

References

Ackerman, T. (1987). Program MANTEL, revised version to incorporate variance estimate of the M-H delta statistic and add a standardization technique.

Allen, M.J. & Yen, W.M. (1979):Introduction to measurement theory, Monterey, California: Brooks/Cole Publishing Company.

Angoff, W.H. & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10 (2), 95-106.

Barcikowski, R.S. (Ed) (1983). Computer packages and research design, Vol. 1: BMDP, Lanham, Md.: University Press of America.

Birch, M W. (1964). The detection of partial association, I: the 2x2 case. Journal of the Royal Statistical Society, Series B, 26, 313-324.

Breslow, N. (1981). Odds ratio estimators when data are sparse. Biometrika, 68, 73-84.

Carlson, J. (1983). Program DATAGEN, IBM version of DATAGEN converted by J. Carlson of the University of Ottawa.

Camilli, G. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Cox, D.R. (1970). Analysis of binary data. London: Methuen and Co, Ltd.

Dizinno, G.A. & Arrasmith, D.G. (1988). Relationship between cultural review of test items and the Mantel-Haenszel statistic. Paper presented at the American Educational Research Association annual meeting, New Orleans, La. April, 1988.

Dorans, N.J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance of the Scholastic Aptitude Test. Journal of Educational Measurement, 23 (4), 355-368.

Flanders, W.D. (1985). A new variance estimator for the Mantel-Haenszel odds ratio. Biometrics, 41, 637-642.

Goddard, R.H. & Lindquist, E.F. (1940). An empirical study of the effect of heterogeneous within-groups variance upon certain F-tests of significance in analysis of variance. Psychometrika, 5 (4), 263-274.

Hambleton, R.K. & Rogers, H.J. (1988). Detecting biased test items: Comparison of the IRT area and Mantel- Haenszel methods. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., April, 1988.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hauck, W.H. (1979). The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. Biometrics, 35, 817-819.

Holland, P.W. (1985). On the study of differential item performance without IRT. Paper presented at the Military Testing Conference, San Diego, September, 1985.

Holland, P.W. & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Educational Testing Service, Program Statistics Research Technical Report no 86-69, Princeton, N.J.

Ironson, G.H. & Craig, R. (1982). Item bias techniques when amount of bias is varied and score differences between groups are present. NIE grant no.G-81-0045, University of South Florida.

Linacre, J.M. (1988). The practical realization of the standard error of the Mantel-Haenszel statistic. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La. April, 1988.

Lindquist, E.F. (1956). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin Company, Boston.

Linn, R.L. & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18 (2), 109-118.

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

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

Martois, J.S., Rickard, P.L., & Stiles, R.L. (1988). Use of the Mantel-Haenszel statistic with test data from workfair participants. Paper presented at a joint session of AERA-NCME annual meeting, New Orleans, La.

Miller, S.K., Doolittle, A.E., & Ackerman, T.A. (1988). Differential item performance for Mexican-American ESL students and white non-ESL students on mathematics and english achievement tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 1988.

Phillips, A., & Holland, P.W. (1986). A new estimator of the variance of the Mantel-Haenszel log-odds-ratio estimator. Educational Testing Service, program statistics research technical report No. 86-67, research report No. 86-21.

Phillips, S.E., & Mehrens, W.A. (1988). Comparison of methods for detecting differential item performance due to instructional/test misalignment. Paper presented at the American Educational Research Association annual meeting, New Orleans, April, 1988.

Plake, B.S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation procedure. Educational and Psychological Measurement, 40, 397-411.

Rudner, L.M. (1977). An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation. Catholic University of America.

Schmeiser, C.B. (1982). Use of experimental design in statistical item bias studies. In Berk, R.A. (Ed). Handbook of Methods for Detecting Test Bias. Baltimore and London: The Johns Hopkins University Press.

Seong, T-J, & Subkoviak, M.J. (1987). A comparative study of recently proposed item bias detection methods. Paper presented at the annual meeting of the National Council of Measurement in Education, Washington, DC, April, 1987.

Shepard, L.A., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22 (2), 77-105.

Simon, M. (1987). Statistical and subjective bias analyses of translated educational achievement items. Unpublished doctoral thesis, University of Toronto.

Skaggs, G., & Lissitz, R.W. (1988). Consistency of selected item bias indices: Implications of another failure. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La, April 1988.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In Berk, R.A. (Ed). Handbook of Methods for Detecting Test Bias, Baltimore and London: the Johns Hopkins University Press, Baltimore and London.

Welch, C.J., Ackerman, T.A., Doolittle, A.E., & Hurley, J. (1987). An examination of statistical procedures for detecting cross-cultural differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, April 1987, Washington D.C.

Welch, C.J., & Doolittle, A.E. (1988). Gender-based differential item performance in english usage items. Paper presented at the annual meeting of the American Educational Research Association, April, 1988, New Orleans, La.

Wright, D.J. (1988). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education in San Francisco.

APPENDIX A:

POST HOC TESTS

Table 16

Post Hoc Tests on the Two-way Interaction of Sample Size and Discrimination on the Standard Deviation of the M-H Z across the 40 Items.

N = 450/150

Source	df	SS	MS	F	P
Among groups	2	.2076	.1038	91.02	.000*
Between groups	297	.3386	.0011		

Scheffe at .05: critical range = .0034

Group identifications: gr 1 : a = 0.7
gr 2 : a = 1.0
gr 3 : a = 1.3

Group means: gr 1 = .1582 gr 2 = .1830 gr 3 = .2221
All groups are significantly different

N = 900/300

Source	df	SS	MS	F	P
Among groups	2	.0562	.0281	109.55	.000*
Between groups	297	.0762	.0003		

Scheffe at .05: critical range = .0016
Group means: gr 1 = .1091 gr 2 = .1271 gr 3 = .1426
All groups are significantly different

N = 1350/450

Source	df	SS	MS	F	P
Among groups	2	.0484	.0242	147.64	.000*
Between groups	297	.0487	.0002		

Scheffe at .05: critical range = .0013
Group means: gr 1 = .0870 gr 2 = .1015 gr 3 = .1181
All groups are significantly different

Table 16 (continued)

"a" = 0.7

Source	df	SS	MS	F	P
Among groups	2	.2656	.1328	546.37	.000*
Between groups	297	.0722	.0002		

Scheffe at .05: critical range = .0016
Group identification : gr 1 : N = 450/150
gr 2 : N = 900/300
gr 3 : N = 1350/450

Group means: gr 1 = .1582 gr 2 = .1091 gr 3 = .0871
All groups are significantly different

"a" = 1.0

Source	df	SS	MS	F	P
Among groups	2	.3474	.1737	542.87	.000*
Between groups	297	.0950	.0003		

Scheffe at .05: critical range = .0018
Group means: gr 1 = .1830 gr 2 = .1271 gr 3 = .1015
All groups are significantly different

"a" = 1.3

Source	df	SS	MS	F	P
Among groups	2	.5912	.2956	296.28	.000*
Between groups	297	.2963	.0010		

Scheffe at .05: critical range = .0032
Group means: gr 1 = .2221 gr 2 = .1426 gr 3 = .1181
All groups are significantly different

Table 17

Post Hoc Tests of the Three-way Interaction on the Standard Deviation of the M-H Z with Sample Size and Discrimination held Constant

"a" = 0.7 N = 450/150

Source	df	SS	MS	F	P
Among groups	2	.0707	.0353	29.03	.000*
Within groups	297	.3616	.0012		

Scheffe at the .05 level: critical range = .0035
Group identification: gr 1 : "b" values from -2 to -.8
gr 2 : "b" values from -.7 to .6
gr 3 : "b" values from .7 to 2
Group means: gr 1 = .1596 gr 2 = .1353 gr 3 = .1723
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

"a" = 0.7 N = 900/300

Source	df	SS	MS	F	P
Among groups	2	.0321	.0160	26.10	.000*
Within groups	297	.1826	.0006		

Scheffe at .05: critical range = .0025
Group means: gr 1 = .1122 gr 2 = .0926 gr 3 = .1163
Groups 1 and 2 and groups 2 and 3 differ significantly

Groups 1 and 3 do not differ significantly

"a" = 0.7 N = 1350/450

Source	df	SS	MS	F	P
Among groups	2	.0196	.0098	29.77	.000*
Within groups	297	.0978	.0003		

Scheffe at the .05 level: critical range = .0018
Group means: gr 1 = .0895 gr 2 = .0747 gr 3 = .0935
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

Table 17 (continued)

"a" = 1.0 N = 450/150

Source	df	SS	MS	F	P
Among groups	2	.1567	.0783	56.76	.000*
Within groups	297	.4099	.0014		

Scheffe at the .05 level: critical range = .0037
Group means: gr 1 = .2007 gr 2 = .1479 gr 3 = .1904
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

"a" = 1.0 N = 900/300

Source	df	SS	MS	F	P
Among groups	2	.0675	.0338	44.36	.000*
Within groups	297	.2261	.0008		

Scheffe at the .05 level: critical range = .0028
Group means: gr 1 = .1355 gr 2 = .1029 gr 3 = .1339
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

"a" = 1.0 N = 1350/450

Source	df	SS	MS	F	P
Among groups	2	.0407	.0204	43.04	.000*
Within groups	297	.1405	.0005		

Scheffe at the .05 level: critical range = .0022
Group means: gr 1 = .1083 gr 2 = .0831 gr 3 = .1073
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

Table 17 (continued)

"a" = 1.3 N = 450/150

Source	df	SS	MS	F	P
Among groups	2	.4162	.2081	46.66	.000*
Within groups	297	1.3245	.0045		

Scheffe at the .05 level: critical range = .0067
Group means: gr 1 = .2413 gr 2 = .1631 gr 3 = .2429
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

"a" = 1.3 N = 900/300

Source	df	SS	MS	F	P
Among groups	2	.1205	.0603	70.70	.000*
Within groups	297	.2531	.0009		

Scheffe at the .05 level: critical range = .0029
Group means: gr 1 = .1542 gr 2 = .1124 gr 3 = .1556
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

"a" = 1.3 N = 1350/450

Source	df	SS	MS	F	P
Among groups	2	.0972	.0486	81.24	.000*
Within groups	297	.1777	.0006		

Scheffe at the .05 level: critical range = .0024
Group means: gr 1 = .1257 gr 2 = .0905 gr 3 = .1311
Groups 1 and 2 and groups 2 and 3 differ significantly
Groups 1 and 3 do not differ significantly

Table 18

Post Hoc Tests on the Percentiles for Sample Size Effect when the Discrimination is held Constant at 0.7.

$P_{2.5}$

Source	df	SS	MS	F	P
Among group	2	1.4590	.7295	145.78	.000*
Within group	297	1.4863	.0050		

Group identification: gr 1 : N = 450/150
gr 2 : N = 900/300
gr 3 : N = 1350/450
Group means: gr 1 =-.3637 =-.2395 gr 3=-.2000
Scheffe at 0.013: critical range = 0.0071
All groups are significantly different.

P_{5}

Source	df	SS	MS	F	P
Among group	2	.9028	.4514	195.69	.000*
Within group	297	.6851	.0023		

Group means: gr 1= -.2842 gr 2= -.1885 gr 3= -.1537
Scheffe at 0.013: critical range = 0.0048
All groups are significantly different

P_{95}

Source	df	SS	MS	F	P
Among group	2	.7953	.3976	243.27	.000*
Within group	297	.4855	.0016		

Group means: gr 1= .2755 gr 2= .1904 gr 3= .1524
Scheffe at 0.013: critical range = .0040
All groups are significantly different

Table 18 (continued)

$P_{97.5}$

Source	df	SS	MS	F	P
Among group	2	1.3265	.6632	159.45	.000*
Within group	297	1.2353	.0042		

Group means: gr 1= .3460 gr 2= .2400 gr 3= .1858
Scheffe at 0.013: critical range = 0.0064
All groups are significantly different

Table 19

Post Hoc Tests on the Percentiles for the Sample Size Effect when the Discrimination Parameter is Equal to 1.0.

$P_{2.5}$

Source	df	SS	MS	F	P
Among group	2	1.9737	.9869	185.28	.000*
Within group	297	1.5819	.0053		

Group means: gr 1= -.4134 gr 2= -.2832 gr 3= -.2184
Scheffe at 0.013: critical range = .0073
All groups are significantly different

P_{5}

Source	df	SS	MS	F	P
Among group	2	1.0600	.5300	256.78	.000*
Within group	297	.6130	.0021		

Group means: gr 1= -.3156 gr 2= -.2211 gr 3= -.1724
Scheffe at 0.013: critical range = 0.0045
All groups are significantly different

P_{95}

Source	df	SS	MS	F	P
Among group	2	1.1603	.5802	176.92	.000*
Within group	297	.9739	.0033		

Group means: gr 1= .3313 gr 2= .2222 gr 3= .1847

Scheffe at 0.013: critical range = .0057

All groups are significantly different

Table 19 (continued)

$P_{97.5}$

Source	df	SS	MS	F	P
Among group	2	1.9188	.9594	127.69	.000*
Within group	297	2.2316	.0075		

Group means: gr 1= .4201 gr 2= .2855 gr 3= .2295
Scheffe at 0.013: critical range = .0087
All groups are significantly different

Table 20

Post Hoc Tests on the Percentiles for the Sample Size Effect when the Discrimination Parameter is equal to 1.3.

$P_{2.5}$

Source	df	SS	MS	F	P
Among group	2	3.3776	1.6880	136.49	.000*
Within group	297	3.6748	.0124		

Group means: gr 1= -.5165 gr 2= -.3251 gr 3= -.2686
Scheffe at 0.013: critical range = .0111
All groups are significantly different

P_{5}

Source	df	SS	MS	F	P
Among group	2	1.8109	.9054	264.88	.000*
Within group	297	1.0152	.0034		

Group means: gr 1= -.3917 gr 2= -.2522 gr 3= -.2098
Scheffe at 0.013: critical range = 0.0065
All groups are significantly different

P_{95}

Source	df	SS	MS	F	P
Among group	2	1.6080	.8040	193.20	.000*
Within group	297	1.2359	.0042		

Group means: gr 1 = .3799 gr 2 = .2534 gr 3 = .2065
Scheffe at 0.013: critical range = 0.0065
All groups are significantly different

Table 20 (continued)

$P_{97.5}$

Source	df	SS	MS	F	P
Among group	2	4.1749	2.0875	48.13	.000*
Within group	297	12.8821	.0434		

Group means: gr 1= .5403 gr 2= .3268 gr 3= .2650
Scheffe at 0.013: critical range = 0.0208
Groups 1 and 2 differ significantly
Groups 1 and 3 differ significantly
Groups 2 and 3 do not differ significantly

Table 21

Post Hoc Test on the Percentiles of the Discrimination Effect when Sample Size is Equal to 450/150.

$P_{.5}$

Source	df	SS	MS	F	P
Among group	2	1.2151	.6076	43.38	.000*
Within group	297	4.1597	.0140		

Group identification: gr 1 : a = 0.7
gr 2 : a = 1.0
gr 3 : a = 1.3
Group means: gr 1= -.3637 gr 2= -.4134 gr 3= -.5165
Scheffe at 0.013: critical range = 0.0118
All groups are significantly different

$P_{.5}$

Source	df	SS	MS	F	P
Among group	2	.6109	.3055	67.70	.000*
Within group	297	1.3401	.0045		

Group means: gr 1= -.2842 gr 2= -.3156 gr 3= -.3917
Scheffe at 0.013: critical range = 0.0067
All groups are significantly different

$P_{.95}$

Source	df	SS	MS	F	P
Among group	2	.5453	.2727	50.31	.000*
Within group	297	1.6095	.0054		

Group means: gr 1= .2755 gr 2= .3313 gr 3= .3799
Scheffe at 0.013: critical range = 0.0074
All groups are significantly different

Table 21 (continued)

$P_{97.5}$

Source	df	SS	MS	F	P
Among group	2	1.9248	.9624	20.64	.000*
Within group	297	13.8491	.0466		

Group means: gr 1= .3460 gr 2= .4201 gr 3= .5403
Scheffe at 0.013: critical range = 0.0216
Groups 1 and 3 differ significantly
Groups 2 and 3 differ significantly
Groups 1 and 2 do not differ significantly

Table 22

Post Hoc Tests on the Percentiles of the Discrimination Effect when Sample Size Equalled 900/300.

$P_{2.5}$

Source	df	SS	MS	F	P
Among group	2	.3658	.1829	32.37	.000*
Within group	297	1.6782	.0057		

Group means: gr 1= -.2395 gr 2= -.2832 gr 3= -.3251
Scheffe at 0.013: critical range = 0.0075
All groups are significantly different

P_{5}

Source	df	SS	MS	F	P
Among group	2	.2029	.1015	51.44	.000*
Within group	297	.5858	.0020		

Group means: gr 1= -.1885 gr 2=-.2211 gr 3= -.2522
Scheffe at 0.013: critical range = 0.0044
All groups are significantly different

P_{95}

Source	df	SS	MS	F	P
Among group	2	.1986	.0993	44.01	.000*
Within group	297	.6703	.0023		

Group means: gr 1= .1904 gr 2= .2221 gr 3= .2534
Scheffe at 0.013: critical range = 0.0048
All groups are significantly different

Table 22 (continued)

$P_{97.5}$

Source	df	SS	MS	F	P
Among group	2	.3771	.1885	34.74	.000*
Within group	297	1.6118	.0054		

Group means: gr 1= .2400 gr 2= .2855 gr 3= .3268
Scheffe at 0.013: critical range = 0.0074
All groups are significantly different

Table 23

Post Hoc Tests on the Percentiles of the Discrimination Effect with the Sample Size held Constant at 1350/450.

$P_{2.5}$

Source	df	SS	MS	F	P
Among group	2	.2519	.1259	41.33	.000*
Within group	297	.9050	.0030		

Group means: gr 1= -.2000 gr 2= -.2184 gr 3= -.2686
Scheffe at 0.013: critical range = 0.0055
Groups 1 and 3 differ significantly
Groups 2 and 3 differ significantly
Groups 1 and 2 do not differ significantly

P_{5}

Source	df	SS	MS	F	P
Among group	2	.1585	.0793	60.76	.000*
Within group	297	.3875	.0013		

Group means: gr 1= -.1547 gr 2= -.1724 gr 3= -.2098
Scheffe at 0.013: critical range = 0.0036
All groups are significantly different

P_{95}

Source	df	SS	MS	F	P
Among group	2	.1485	.0742	53.07	.000*
Within group	297	.4155	.0014		

Group means: gr 1= .1524 gr 2= .1847 gr 3= .2065
Scheffe at 0.013: critical range = 0.0037
All groups are significantly different

Table 23 (continued)

$\underline{P}_{97.5}$

Source	df	SS	MS	F	P
Among group	2	.3142	.1571	52.53	.000*
Within group	297	.8882	.0030		

Group means: gr 1 = .1858 gr 2 = .2295 gr 3 = .2650
Scheffe at 0.013: critical range = 0.0055
All groups are significantly different