INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9224863

Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning

Uttaro, Thomas Edward, Ph.D.

City University of New York, 1992

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A

FACTORS INFLUENCING THE MANTEL-HAENSZEL PROCEDURE IN THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING

by

Thomas E. Uttaro

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1992

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

April 6, 1992
Date

Roger E. Millsap
Chair of Examining Committee

April 6, 1992
Date

Executive Officer

Professor Alan Gross

Professor Howard Everson

Professor David Rindskopf

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

FACTORS INFLUENCING THE MANTEL-HAENSZEL PROCEDURE IN THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING

by

Thomas E. Uttaro

Advisor: Professor Roger Millsap

A continuing inquiry has focused on the development of statistical procedures for DIF analysis which require smaller sample sizes and are more practical and less costly than IRT based procedures, (Lord, 1980; Rudner, Getson & Knight, 1980; Linn & Harnish, 1981). The application of the Mantel Haenszel procedure to DIF study (Mantel & Haenszel, 1959; Holland & Thayer, 1988) is the most recent of continuing efforts to cultivate chi-square methodology. The Mantel-Haenszel procedure is intended to approximate the theoretically preferred IRT definition of DIF (Lord, 1980). Correspondence has been reported between the MH procedure and IRT methods in several DIF comparison studies (Thissen, Steinberg & Wainer, 1988; Hambleton & Rogers, 1989).

Simulated data sets were generated from IRT DIF models and the MH procedure was applied to recover the IRT DIF. The MH procedure was evaluated in terms of the statistical accuracy of chi-square type I and type II error rates, α_{mh} and Δ_{mh}

estimators, variances and ETS DIF classifications. The results of the study suggest that there are serious obstacles to the simple, direct application of the Mantel-Haenszel procedure for DIF analyses.

The procedure may lead to inaccuracies given certain item ICC parameterizations of studied items and items in the matching criteria and given discrepancies between focal and reference group Θ levels. In the Rasch case the expected behavior of α is verified but instability of MH_{CHISQ} is demonstrated under discrepant ability and matching criterion item difficulty conditions. Factors influencing the empirical and estimated (Phillips & Holland, 1987) variances of α are examined. It is shown that the estimated variance is sensitive to ICC parameterizations and ability discrepancies. The reliability of total score given known ability (Θ) estimates is evaluated and shown to interact with ability discrepancies, influencing the MH procedure.

Acknowledgements

Foremost, I thank my wife, Karen Uttaro, whose love, patience and understanding made it possible for me to complete my doctoral studies. She has sustained me throughout this endeavor, providing hope and encouragement. She has fulfilled my responsibilities when I have been unable to do so. My love and respect for her is enduring.

I am very grateful to my parents, Camille and Ralph Uttaro, who nurtured an appreciation of learning and education in our family. Their emotional and material support has been essential over the course of my educational career.

I am thankful to my brother, James Uttaro, and my sister, Suzanne Samuels, for providing support and perspective throughout this process. I also thank my grandfather, Sebastian Amoruso, for always expressing confidence in me. I appreciate the backing of my extended family and friends who have inquired after my progress.

I am deeply indebted to my mentor, Professor Roger Millsap, who has been very generous in sharing his ideas and expertise with me. He has been responsive to my thoughts and concerns, making the dissertation process enjoyable. I am also indebted to Professors Howard Everson, Alan Gross, Carol Tittle and David Rindskopf for serving on my committee. I thank them for careful reading of this manuscript and their valuable suggestions. I also thank my colleagues at South Beach Psychiatric Center for their ongoing support.

TABLE OF CONTENTS

Approval....ii

Abstract....iii

Acknowledgements....v

Table of Contents....vi

List of Tables....viii

List of Figures....x

I. Introduction....1

II. Literature Review

A) Classical DIF approaches....11

B) Observed and Unobserved Conditional Invariance.13

C) Chi Square DIF Approaches....14

D) Item Response Theory Approaches

i) Outline of the Model....17

ii) Item Response Theory DIF Methods....22

E) Mantel-Haenszel Statistic....27

F) Mantel-Haenszel DIF Studies....32

G) Summary and Purpose....46

i) Phase I Studied Item ICC Parameter and Ability Factors....46

ii) Phase II Rasch Cases....47

iii) Phase III Variances....48

iv) Phase IV Criterion Items-Eta Analyses....49

III. Method....52

A) Phase I....53

B) Phase II....55

C) Phase III.....56

D) Phase IV.....57

IV. Results

A) Phase I.....62

B) Phase II.....86

C) Phase III.....89

D) Phase IV.....98

V. Discussion.....105

VI. Summary.....112

Appendix A.....114

References.....116

LIST OF TABLES

Table

1. Pilot Eta Manipulations.....60
2. MH chi square type I error rate, ETS DIF classifications.....63
3. MH chi square type II error rate, ETS DIF classifications.....64
4. MH odds ratio, ETS delta, no DIF modelled.....65
5. MH odds ratio, ETS delta, DIF modelled.....66
6. Variances: Estimated, empirical and discrepancies no DIf modelled.....67
7. Variances: Estimated, empirical and discrepancies.....68
8. ANOVA chi square type I error rate by A, B, C and θ_{dis}.75
9. ANOVA chi square type I error rate by A, B, C and θ_{dis} summary means A by C interaction.....76
10. ANOVA chi square type I error rate by A, B, C and θ_{dis} summary means A by C by θ_{dis} interaction.....77
11. ANOVA α by A, B, C and θ_{dis} type I conditions.....78
12. ANOVA α by A, B, C and θ_{dis} type I conditions summary means A by C by θ_{dis} interaction.....79
13. ANOVA chi square type II error rate by A, B, C and θ_{dis}82
14. ANOVA α by A, B, C and θ_{dis} type II conditions.....84
15. Rasch Cases.....88
16. ANOVA empirical variance by A, B, C and θ_{dis} type I conditions.....91

17. ANOVA estimated variance by A, B, C and θ_{dis} type I conditions.....92
18. ANOVA empirical variance by A, B, C and θ_{dis} type II conditions.....93
19. ANOVA estimated variance by A, B, C and θ_{dis} type II conditions.....94
20. ANOVA Variance Discrepancy by A, B, C and θ_{dis} type I conditions.....96
21. ANOVA Variance Discrepancy by A, B, C and θ_{dis} type II conditions.....97
22. Eta^2 for reference and focal groups, chi square type I error rate, ETS DIF classifications.....99
23. Eta^2 for reference and focal groups, chi square type II error rate, ETS DIF classifications.....100
24. Eta^2 multiple regression type I conditions.....101
25. Eta^2 multiple regression type II conditions.....102

LIST OF FIGURES

Figure

1. MH11/MH17.....69
2. MH12/MH18.....69
3. MH13/MH19.....69
4. MH14/MH110.....69
5. MH15/MH111.....70
6. MH16/MH112.....70
7. MH21/MH27.....70
8. MH22/MH28.....70
9. MH22R/MH28R.....71
10. MH23/MH29.....71
11. MH23R/MH29R.....71
12. MH24/MH210.....71
13. MH25/MH211.....72
14. MH26/MH212.....72
15. MH31/MH37.....72
16. MH32/MH38.....72
17. MH33/MH39.....73
18. MH34/MH310.....73
19. MH35/MH311.....73
20. MH36/MH312.....73

I. Introduction

In recent years a plethora of methods have been developed for the detection of item bias or differential item functioning (DIF). The Mantel-Haenszel procedure has been proposed as an accurate, inexpensive and easy to apply DIF technique, (Holland and Thayer, 1988). It is intended to approximate item response theory DIF procedures (Rudner, 1977; Lord, 1977, 1980; Linn & Harnish, 1981) which are based on the "theoretically preferred" IRT DIF definition of the equivalence of item characteristic curves. The research reported in this paper evaluates statistical aspects of the MH procedure as it reflects IRT DIF modelled conditions in simulated data.

Items should function similarly in all subgroups in the sense that the relationship between scores and ability should be the same for the subgroups. If this relationship is not invariant, undesirable factors are influencing item performance and DIF exists. In situations in which DIF affects item performance between two demographic groups, a serious threat to the validity of the test exists. The implications and conclusions drawn from the test are likely to be flawed and the selection consequences are likely to be detrimental to interests of minorities and women. Access to educational and economic opportunities often depends on individuals' ranking on measurement instruments. If we are to apply scientific principles to educational and psychological

measurement we must insure that tests really measure what they are intended to and that their results are not the outcome of extraneous factors such as group membership.

In the development of tests there should be careful consideration given to the degree of familiarity in terms of substantive content and the presentation and meaning of the item for examinee subgroups. It is quite possible that subgroups have not had equal opportunities to become familiar with substantive content. There is little doubt that ability levels are strongly influenced by the quality of educational opportunities afforded to examinee subgroups. This condition alone does not invalidate the constituent items or the test as a whole, however, even though it indicates an undesirable disparity in terms of the access of subgroups to balanced educational opportunities. Rather it is the performance of subgroups after accounting for ability that indicates potential bias of an item. Subgroup members should have an equal chance of succeeding on an item with ability 'partialed out'. To the extent that this is not so, membership in a subpopulation group will influence item performance. DIF procedures are based on statistical models which are intended to distinguish the separation of persons on the basis of ability from separation based on extraneous factors. The measure of ability is a critical issue in DIF procedures the nature of which distinguishes families of procedures (Millsap & Meredith, in press).

Tests are intended to sample latent variables of particular relevant interest in terms of a curriculum or other established paradigm. Differences between persons should emanate from these relevant sources only. Items showing DIF do not properly sample the latent variable of interest and are affected by inadvertent linkages to cultural factors. Tests are frequently constructed to measure performance criteria associated with the culture of the American middle-class. Some theorists have argued that these relate best to job and school success (Lyman 1986), and therefore represent relevant criteria. However, if tests too closely parallel white, middle class culture they may alarmingly underestimate the ability of minorities and women through inappropriate sensitivity to group differences. Statistical approaches to DIF or item bias are not intended to determine the relevance of particular criteria. That issue is properly the focus of a larger educational debate on policy, programs and curricula. However, statistical approaches assume that performance relevant to criteria can be measured independently of the influences of group membership.

There is a clear consensus that testing and evaluation programs must work to reduce influences based on racial, ethnic, gender, socioeconomic and regional specificity. The practice of testing should be consistent with democratic values. Performance on measurement instruments should not be influenced by specific group membership of examinees beyond

the extent to which group membership happens to be correlated with ability. Certain groups may have advantages that are closely related to socioeconomic and educational privileges resulting in higher overall levels of the ability being examined. Testing procedures cannot equalize disparities in abilities resulting from such advantages. Nor would this be desirable if tests are to accurately measure abilities which may be distributed unevenly in the population across certain groups.

The statistical analysis of individual test items to determine the potential for bias is a recent approach in modern testing. These procedures have only been developed over the last two decades and their validity and utility is not entirely known. It is important then to consider whether various statistical procedures have the power to clearly differentiate items in which DIF may be operating. This capacity is derived from the total information available in the test, based on all the items, which provides the basis for ascertaining the potentially aberrant behavior of a particular item. Thus, the total test is regarded as less biased than a particular item and an item is considered biased in relation to the total test (Ironson, 1982). An item is considered unbiased if individuals with equal ability but from different groups have the same probability of answering the item correctly (Lord, 1980). Thus, procedures which can effectively control for ability and reasonably model the data

should be best for bias analyses.

Since tests are used for selection there is typically an underlying belief that by removing items which are internally problematic and potentially biased the likelihood of selection bias will be reduced. However, some recent work has shown that a test can be free of biased items and yet result in selection bias, (Millsap & Meredith, in press). The ideal outcome is that inferences drawn from the test should be valid by both internal and external criteria.

Although selection bias and item bias are considered distinct concepts they have similar heuristic foundations. Raju & Normand (1985) showed that a unitary regression model can be applied to evaluate selection bias and item bias. A method for predicting success on an item from total score is developed. By examining the fit of the regression line for two groups, inferences can be made about potential item bias. Such a unified approach suggests linkages between internal and external validity, since differential item functioning may have selection consequences.

Large testing organizations such as the Educational Testing Service and the American College Testing Program, Inc. devote considerable resources to developing procedures for detecting DIF. Due to large cohorts of examinees year after year and the ability of such organizations to apply the most sophisticated DIF procedures, such as those based on IRT, we might expect considerable success in their application of DIF

procedures. We note, however, the recent court findings of gender based bias in the Scholastic Aptitude Test, administered by the College Board. In Sharif vs. New York State Education Department, (1989), the United States District Court for Southern New York issued an injunction preventing the State Of New York from awarding Regents College Scholarships for the 1988 academic year based on SAT scores alone. The plaintiff successfully argued that females were not receiving scholarships in the proportion expected from their high school performance and that this was a function of pervasive gender bias in the SAT. Since the SAT is a wide range achievement (ability) test there should be a close correspondence between high school performance and SAT scores. Females, although having slightly higher levels of high school performance, were not scoring as expected on the SAT and therefore not receiving Regent's Scholarships at the expected rate. Although this case was not decided on bias in specific items, there is strong suspicion of item bias on the quantitative portion of the test. DIF detection procedures had not picked up the apparent bias nor provided evidence otherwise. The State of New York and the College Board have chosen not to return to court to challenge the preliminary injunction.

A consensus appears to have been reached that the theoretically preferred approach to testing DIF is that based on item response theory because of the sample invariant

properties of item parameters when an IRT model fits the data. Several comparison studies have endorsed IRT methods and asserted that various chi-square techniques are the best alternatives (Ironson & Subkoviak, 1979; Rudner et al 1980; Subkoviak et al, 1984). Substantive reviews have recommended the use of IRT procedures because of overall accuracy (Ironson, 1982) and because they are most successful at eliminating differences in ability as a confounding factor between groups (Burrill, 1982). The power of IRT DIF methods is a result of the conditioning on ability. This permits sample invariant item parameters to be estimated for each group which can then be compared in DIF analyses. However, to obtain stable parameters it is generally necessary to perform calibration on large samples for both the focal and reference groups except under the Rasch or one parameter logistic model. Swaminathan and Gifford (1979), in a simulation study, obtained a correlation of .88 between known theta (θ), the ability parameter under IRT models, and predicted θ for 1000 simulated examinees under a three parameter model. For smaller subject groups the correlation fell off sharply. Lord (1968) indicated that approximately 1000 subjects are needed to obtain stable discrimination (a) parameters under three parameter IRT models. Under the Rasch model stable parameter estimates can be obtained with 250 subjects (Wright & Stone, 1979). A drawback of IRT based methods is that it is necessary to utilize a sophisticated calibration program such

as LOGIST (Wood, Wingersky, & Lord, 1976) or BILOG (Mislevy & Bock, 1982). These require support by advanced computer hardware systems.

Thus research continues in an effort to develop statistical procedures for DIF analysis which require smaller sample sizes, are more practical and less costly. Linn & Harnish (1981) advanced a pseudo-IRT method which attempts to preserve the power of the procedures but is applicable with much smaller sample sizes. Support was found for the use of pseudo-IRT and chi-square approaches in Shepard, Camilli and Williams (1985) who compared several approximation techniques to an IRT based method.

Chi-square methods for examining DIF have a straightforward statistical foundation and are computationally simple. The application of the Mantel Haenszel procedure (Mantel & Haenszel, 1959) to DIF study is the most recent of continuing efforts to cultivate chi-square methodology (Holland & Thayer, 1988). A close correspondence has been reported between the MH procedure and IRT methods in several DIF comparison studies. Hambleton & Rogers (1989) report consistent classifications of items into biased and unbiased categories leading to similar results for both methods. Thissen, Steinberg and Wainer (1988) reported "striking" agreement between the MH and IRT likelihood ratio procedures. Theoretically, the IRT likelihood ratio procedure should show more power than the MH procedure when the IRT model fits,

since it directly tests the exact difference in examinee data. They found no such disparity, and concluded that MH was just as likely to produce chi-squares indicating DIF as the IRT likelihood ratio procedure. These analyses indicate that if MH is applied appropriately it should produce results that are precise, accurate and closely correspond to item response theory based techniques.

It might be expected that the need for accurate DIF procedures is even more acute in smaller organizations or those not exclusively devoted to testing. Elementary and secondary schools, colleges, universities and employers all have need to practice DIF free measurement and may not have simple, thoroughgoing procedures for doing so. The promise of the Mantel Haenszel procedure is that it may provide just such a simple, useful and accurate measure of DIF. It is easily applied, requires low computing costs and is an index with an understandable metric. It also has desirable statistical properties in that the MH odds ratio estimates have a known interpretation and that the MH statistic is a significance test.

The performance of the MH statistic has yet to be thoroughly assessed and more evidence is required to substantiate its widespread utility. There are particular conditions in terms of item response function parameters, ability distributions of reference and focal groups, and reliability of total scores and the underlying latent trait

that motivated this study. The study reported here attempts to contribute to what is known about the accuracy of the MH statistics, especially the MH chi-square test, under certain conditions. The following literature review is intended to demonstrate the need for studying these conditions in the context of ongoing DIF research.

II. Literature Review

The literature review is intended to set this research in an historical and technical context. It is comprised of eight sections leading to the specific questions being addressed by this research project.

A. Classical Item Bias Approaches

Hulin, Drasgow & Parsons (1983) reviewed much of the item bias research motivated by classical test theory. These approaches focus on the p_i values for different subgroups, where p_i is the proportion of a sample responding correctly to an item. When the set of items is unbiased the relative item difficulties or rank order of item difficulties should not differ markedly across two subpopulations (Angoff & Ford, 1973). These values are plotted against each other, a best fitting line is considered, and outliers are regarded as possibly biased. Problems in terms of the non-linearity of the plot of p_i values and the finding that p_i values are not normally distributed are dealt with by plotting a standard transformation of them. ETS delta values (Donlon & Angoff, 1971) transform p values via $\Delta_i=4\Phi^{-1}(1-p_i)+13$, where Φ^{-1} is the inverse normal transformation that converts a probability value into a normal deviate with unit variance. Thus, the transformed p value is placed on a normal curve with a mean 13.0 and standard deviation 4.0. The area under this normal curve above the item is the proportion of persons passing the item among those who reach it. If the distribution of p

values is spread evenly between 0 and 1 the Δ values will be normally distributed. This method assumes that the differences in ability distributions between groups will affect the item difficulties in a similar manner, justifying a best fitting line approach. However, Lord (1980) shows that if subgroup item characteristic curves derived from the p_i values differ on the discrimination parameter ("a" parameter in an IRT model) artifactual bias may be found between the items.

Cleary and Hilton (1968) proposed an approach for detecting biased items which evaluates item by test score interaction. Relative performance of each group on particular items and the total test is examined. Bias is inferred if the patterns of item and test performance are different between the groups. An example of such a disturbance would be if the relationship of item and test difficulties (p-values) was reversed between the two groups.

In another classical test theory approach, item discrimination statistics, such as biserial or point biserial correlations between items and total test scores, are evaluated. The comparison of these correlations across samples is used to infer the extent to which responses to studied items are related to the trait measured by the other items comprising the test. If the items are unbiased, the item-total correlations should be equivalent across groups for various ability levels (Green & Draper, 1972). Items should

discriminate equally well among ability levels for all groups if they are not biased. Otherwise, the item may contain certain information available for one group and not the other and therefore may be regarded as potentially biased. Correlations are sensitive to restriction of range and differences in difficulties and variances of items. Therefore, differences in ability distributions between the groups will influence item discrimination statistics making their use in item bias studies problematic.

B. Observed and Unobserved Conditional Invariance

Modern DIF methods require that examinees be matched in some way on ability so that item by group, rather than item by group by ability, interactions can be evaluated. Thereby only comparable members of the focal and reference groups are compared. Two broad approaches are distinguished by the general methods employed for the control of ability (Millsap & Meredith, in press; Meredith & Millsap, in press). Examinee ability can be defined in terms of an observed measure such as total test score or by a probabilistic function relating examinee ability and item response function parameters to item performance. In the first approach, observed conditional invariance (OCI) is estimated. An item is unbiased if, after matching on an observed ability measure such as total test score, item scores are unrelated to demographics.

In the second or unobserved conditional invariance (UCI) approach the differential probability of success for each

group is assessed conditional on a latent trait. IRT based methods for detecting DIF evaluate UCI through the equivalence of item response functions which condition on θ, the latent ability trait.

An item is considered unbiased if the item response functions are identical among demographic groups. These relations can be summarized as follows: let Y=item score, W=measure on the latent trait, Z=observed ability measure (such as total score) and V=demographic variable. UCI holds if and only if

$$P(Y|W=w, V=v) = P(Y|W=w) \tag{1}$$

and OCI holds if and only if

$$P(Y|Z=z, V=v) = P(Y|Z=z) \tag{2}$$

for all W, V and Z. A major limitation of bias detection methods that evaluate (2) is that OCI need not imply, or be implied by UCI. In fact, such implications hold only under a limited set of conditions. Thus, methods which test OCI do not necessarily test UCI.

C. Chi-Square Approaches

Chi-square approaches clearly fall under the OCI framework as total score or discrete intervals based on total score define the ability continuum. A chi-square approach for evaluating DIF utilizing only correct responses is advanced in Scheuneman (1979). Subjects are grouped into total score

categories and a chi-square is computed through:

$$X^2_{correct}=\sum \frac{(E_{ij}-O_{ij})^2}{E_{ij}} \quad (3)$$

using only correct responses where j=score group, i=focal/reference group, E=expected cell frequency correct, O=observed cell frequency correct. One X^2 is computed for each score group and subject group and the statistic is evaluated with (K-1) degrees of freedom.

Because the Scheuneman procedure uses only the correct responses it has been found not to be satisfactorily distributed as a chi-square statistic (Marascuilo & Slaughter 1981). Further, it will not reject the null hypothesis satisfactorily, resulting in excess type II errors.

A full chi-square method utilizing expected frequencies for correct and incorrect responses was proposed by Shepard, Camilli & Averil (1981). It is essentially the sum of the chi-squares for the correct and incorrect responses:

$$X^2_{full}=X^2_{correct}+X^2_{incorrect} \quad (4)$$

It can also be conceptualized as in the following table for each ability level, j=1...K.

Right/Wrong by Reference/Focal group

Subgroup	Right (1)	Wrong (0)	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{fj}
Combined	M_{1j}	M_{0j}	T

A, B, C and D are frequency counts for the respective cells, each of the score levels contributes to the overall chi-square through the relation

$$X^2_{full}=X^2_1+X^2_2\ldots X^2_k. \tag{5}$$

Chi-square is calculated through

$$X^2_j=\frac{N_j(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}. \tag{6}$$

Ironson (1980) demonstrates that these equations are sensitive to sample size and expected cell frequencies. Cells at the extremes of the ability distributions may have small frequencies for reference or focal groups because of differential overall ability. In such a situation the presence of cells having small frequencies may cause departures from the chi-square distribution in the overall statistic. The location of boundaries of ability groups also affects the level of the chi-square statistic. Variations of

the boundary locations can influence the chi-square to be significant or non-significant.

D. Item Response Theory Approaches

i) Outline of the model

There are several IRT based item bias detection techniques which evaluate a particular set of item response parameters between the focal and reference groups for differences. Item charactistic curves or item response (characteristic) functions describe the probability of a correct response over different levels of a latent ability trait Θ. The value of P(x) is specified in the general form

$$P(X_k) = \int_{\theta} \prod_{k=1}^{J} P_k(\theta)^{X_k} Q_k(\theta)^{1-X_k} \, dG(\theta) \qquad (7)$$

which is the proportion of the population of examinees who would produce the observed response vector if tested. G(Θ) represents the distribution function of latent ability across the population of examinees. $P_k(\theta)$ is the item characteristic curve (ICC) for item k. These ICCs are typically monotonically increasing functions in which examinees with higher levels of the latent trait will have a higher probability of a correct response. A one, two or three parameter model is specified in which discrimination, difficulty and pseudo-guessing parameters may be present in the function. These functions may be based on normal or logistic ogives. The latter are often favored because of

their ease of computation. If a scaling factor of D=1.7 is introduced, the normal and logistic ogives do not deviate by more than .01 on any point of the θ scale. Models may be described by use of either ogive and there is a corresponding normal model for each of the logistic models. For simplicity only the logistic models are described here.

In IRT models we typically assume that examinee response probabilities are a function of θ, a single latent characteristic of the individual. The latent trait space is therefore considered unidimensional in that no other factor is thought to influence the probability of response patterns. In theory, multidimensional models are permitted.

Conditional or local independence of items is a consequence of a unidimensional latent space. Conditional independence implies that among examinees with common trait or ability values, item performances are independent across items. If this condition holds then the probability of any particular response vector is simply the product of the probability of occurrence of scores for each test item. Responses given by an examinee are independent, conditional upon θ. A consequence of local independence is that the likelihood function for a particular response vector can be evaluated. If X_i i=1...p are dichotomous item scores, $P_i(\theta)=P(X_i=1|\theta)$ and $Q_i=1-P_i(\theta)$ then

$$L=\prod_{i=1}^{p} P_i(\theta)^{X_i} Q_i(\theta)^{1-X_i}. \qquad (8)$$

This likelihood is for the vector of responses for a single examinee with the probabilities multiplied together. Conditional or local independence can be obtained whenever the complete latent space has been specified.

The one parameter logistic or Rasch model (Rasch, 1966) takes the form

$$P_i(\theta) = \frac{e^{D(\theta-b)}}{1+e^{D(\theta-b)}}. \qquad (9)$$

where Θ is the level of the latent ability trait, D is the scaling factor of 1.7 to approximate the normal function and b is the index of item difficulty. The b parameter is the point on the ICC of Θ at which the inflection of the slope occurs. In a one or two parameter model it represents the point on the ability scale at which an examinee has a 50% chance of answering the item correctly. Under the one parameter or Rasch model, abilities and difficulties are measured on the same scale and unweighted total score is a sufficient statistic for ability if the model fits (Lord & Novick, 1968; Wright & Stone, 1979). Sufficiency of total score in an IRT model means that all the information available in the response vector regarding Θ is contained in an examinees' total score.

A two parameter IRT model incorporates an "a" or discrimination parameter and takes the logistic form

$$P_i(\theta) = \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}. \qquad (10)$$

The discrimination parameter is proportional to the slope of $P_i(\theta)$ at the point $\theta=b_i$. It is at this value that the slope of the ICC is maximized as the a parameter controls the steepness of the ICC in the region near b. When a is large the ICC is steepest in the region near b and discriminates well there. If the a parameter is small the ICC will rise smoothly over a wide interval centered at b_i.

Items with smaller a values discriminate less well than items with large a values near their corresponding b values. They have more discriminating power at values relatively far from b. An index of dispersion 1/a indicates how much horizontal spread there is in the ICC. Larger values of a will result in smaller values for 1/a and less spread in the ICC.

In the full three parameter logistic model a pseudo-guessing or c parameter is incorporated. It is the probability of a correct response for very low ability persons. The model takes the logistic form:

$$P_i(\theta) = c + (1-c)\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} \qquad (11)$$

c is a lower asymptote.

Parameters are usually estimated through maximum likelihood or marginal maximum likelihood procedures. Hambleton & Swaminathan (1985) maintain that maximum likelihood estimators possess several desirable properties including consistency (ML estimators asymptotically converge

to true values), efficiency (ML estimators asymptotically have the smallest variance), sufficiency when such statistics exist and an asymptotic normal distribution.

The asymptotic sampling variance of the maximum likelihood ability estimate is the reciprocal of the information function evaluated at ability θ. This is a major feature of IRT models, and is expressed as follows:

$$I_i(\theta) = -E[\frac{D^2 \ln L}{d\theta^2}] = \Sigma \frac{\acute{P}^2}{P_i Q_i}, \qquad (12)$$

E[ ] is an expectation and L is the likelihood function previously defined in (8). Maximum information is attained at the point b_i in a one or two parameter model. The statistic $1/\sqrt{I(\theta)}$ is the standard error of estimate of predicted θ.

The item information function is used to assess the level of confidence in the estimate of θ. The relative degree of item information across the θ range affords an evaluation of the certainty of θ estimates. Items typically provide more information regarding θ estimates at certain θ ranges than others. The amount of item information is a function of the steepness of the ICC and the conditional variance of parameter estimates at various levels of the θ continuum.

The test information function is the sum of all the item information functions and thus is given by:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (13)$$

It gives an indication of amount of information the overall test provides regarding θ estimates at particular levels of the continuum. The test information function may be considered the IRT measure analogous to test score reliability since it sums the item information functions. High test information enhances replication of the ordering of examinees as an outcome of the test. A test having high levels of information across the range of θ would be reliable across that range in the classical sense. Information made available by a test at a particular θ is inversely related to the accuracy of the θ estimate at that point through the standard error of estimation:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}. \tag{14}$$

This provides for a confidence band around θ estimates. The information functions have important applications in the areas of item selection, test construction, and test comparison.

ii) Item Response Theory DIF Methods

According to Lord (1980) the hypothesis of no dif in an IRT model in item k is Ho(IRT):

$$P_{kr}=(\theta)=P_{kf}(\theta)=P_{k}(\theta), \tag{15}$$

that is, the ICCs are equivalent for both groups.

IRT DIF methods are frequently endorsed because parameter estimation is independent of distribution of abilities and the latent trait is considered a more pure description of

"ability" than such a measure as total score. Therefore, these approaches are based on evaluating unobserved conditional invariance (UCI) previously described.

Rasch model procedures for detecting DIF follow two approaches assuming that calibrations have been performed and the Rasch model fits. The equality of difficulty levels of an item across groups is investigated in Wright, Mead & Draba (1976) by means of the statistic:

$$Z=\frac{b_r-b_f}{[SE(b_r)+SE(b_f)]} \tag{16}$$

which is distributed asymptotically N(0,1).

Wright & Stone (1988) outline a procedure for graphically identifying potentially biased items. Separate calibrations are performed for the R and F groups and respective item difficulties are plotted against each other. A line is put through the origin with a slope of 1 and 95% control lines (confidence interval lines) are drawn. Items which fall outside the control lines are considered potentially biased. A test of fit can also be examined, (Wright & Stone 1979). If an item has similar difficulties for subgroups but significant misfit as indicated by infit and outfit statistics, bias may be indicated.

Lord (1980) developed a method for testing the equivalence of the ICC in the focal and reference groups based on the equality of parameter values. He showed that one parameter (difficulty) can be tested by comparing the b estimates with

a statistic equivalent to the Wright, Mead & Draba (1976) method. Lord also developed the IRT-D^2 approach demonstrating that asymptotic normal theory provides for a simultaneous test of differences of the a and b parameters by evaluating the Mahalanobis distance between two vectors. To test such equality the following is calculated:

$$D^2 = \mathbf{v}' \Sigma^{-1} \mathbf{v} \tag{17}$$

v is the vector of the differences between parameter values in the reference and focal groups $[a_r-a_f, b_r-b_f]$ and Σ is their estimated covariance matrix. D^2 is distributed as a Chi-Square statistic with 2 degrees of freedom.

IRT-LR procedure uses likelihood ratio chi squares to test parameter equivalence. This approach assumes that the chosen IRT model fits well in each group and is outlined in Thissen, Steinberg & Wainer (1988) for the k=1 parameter case ($b_r=b_f$). The k≠1 parameter case is a simple generalization. Calibrations are performed to fit the model to both groups simultaneously with certain anchor items (other than the studied item) having common b parameter values. For the studied item the b parameter values differ ($b_r \neq b_f$). X_1^2=-2 log(likelihood) is calculated for the maximum likelihood parameter estimates. The model is then fit under the restriction $b_r=b_f$ and X_2^2 is calculated. The Neyman-Pearson likelihood ratio test is then performed $X^2=X_2^2-X_1^2$. The Chi-Square is evaluated on k degrees of freedom. This research

shows that the IRT-LR test is the uniformly most powerful available as it is a generalized likelihood ratio test of IRT parameters. As such, it maximizes the rejection region of the null hypothesis that the ICC parameters are identical between the focal and reference groups, minimizing Type II error if the IRT parameterization actually holds. IRT-LR and IRT-D^2 are asymptotically equivalent and depend the multivariate normality of the likelihood of parameter values when obtaining p-values from the chi square distribution. In addition, IRT-D^2 requires an accurate estimation of Σ. To the extent that Σ is not well estimated it may be advantageous to utilize IRT-LR procedure.

Rudner (1977) developed a strategy for examining the eqivalence of item response functions by calculating the area between the estimated ICCs. After calibrating the parameters for the two groups and standardizing on b, the ability scale is divided into small intervals (.005) the statistic

$$A_{1i} = \sum_{\theta=-3}^{+3} |P_r(\theta_k) - P_f(\theta_k)| \Delta\theta \tag{18}$$

is calculated by summing over the intervals. Linn et al (1981) developed a similar statistic

$$A_{2i} = \sum_{\theta=-3}^{+3} \{[P_r(\theta_k) - P_f(\theta_k)]^2 \Delta\theta\}^{\frac{1}{2}} \tag{19}$$

and recommended weighting by the standard error of estimate of $P_i(\theta_k)$ to account for the accuracy of estimation along the θ continuum. Other studies have integrated over either an

infinite or finite ability continuum (Rudner et al 1980).

While area measures between ICCs appear analytically desirable, no direct significance test of area differences is available. Typically a cutoff value is determined by carrying out the analysis on two randomly equivalent groups and assigning the highest area value to be the largest to possibly occur by chance (Hambleton & Rogers 1988). Raju (1988) discussed some difficulties in carrying out area integrations over finite or infinite ability continua. The question arises of the meaning of integrating or taking squared differences over strata of the θ continuum where few examinees are located. The magnitude of area measures also depend partly on the length of the θ continuum over which the area is calculated. For example, when the c parameters are unequal and the integration is carried out over an infinite continuum the area measure will be infinite. In addition, standard error formulas are not currently available to evaluate the stability of area measures as they relate to the quality of parameter estimation. Devine & Raju (1982) found weak intercorrelations between the chi-square method and the area based IRT method. There was little agreement in classifying biased items and the percent agreement between the methods appeared to be related to the extent of bias in the studied items.

Thissen & Wainer (1990) illustrate a potentially useful approach to establishing confidence intervals for IRT models.

These depend on the relative density regions of the item parameter spaces which are treated as posterior densities. In a Bayesian analysis M-lines are plotted which depend on the relative frequencies of parameter vectors. Those near the parameter estimates are considered more likely. The M-line plots approximatly fill the (1-α) confidence interval. It is suggested that confidence intervals could provide an estimation of the sampling variation of ICCs and the degree of their overlap may provide an indication of DIF. Such a procedure might refine IRT-DIF analyses based either the equivalence of parameters or area measures.

E. Mantel-Haenszel Statistic

Holland and Thayer (1988) outline the Mantel-Haenszel procedure as a noniterative contingency table method. The procedure is intended to detect test items that function differently in two groups and is conceptually related to earlier chi-square DIF methods. It meets the modern criteria for DIF analysis in that it contrasts comparable members of the two groups by matching on ability level taken as the total score on the test. Hence, it is an observed conditional invariance (OCI) based method. The method estimates and tests a common two factor association parameter α, the common odds ratio, in K 2 X 2 tables. There are K ability intervals, one for each level of total score or grouped total score intervals. These levels are used to match the samples from each group. For each of K ability levels or intervals a 2 X

2 contingency table is constructed as for the simple chi-square case on pg. 16.

Two statistics are evaluated 1) common odds ratio α_{mh} across the 2 X 2 tables 2) The Chi Square test of the Ho that the odds ratios are equivalent for both groups against the specific alternative hypothesis that they differ

$$H_1: \frac{P_{rj}}{Q_{rj}} = \alpha P \frac{fj}{Q_{fj}} \qquad (20)$$

j=1,...,K (score groups).

The Mantel-Haenszel odds ratio estimator is given as

$$\hat{\alpha}_{mh} = \frac{\Sigma_j A_j D_j / T_j}{\Sigma_j B_j C_j / T_j} \qquad (21)$$

A, B, C and D are the cells of the R/F group by correct/incorrect contingency table on pg. 16, one for each level of K. α is the average factor by which the odds that a member of the reference group is correct on the studied item exceeds the corresponding odds for a comparable member of the focal group. The significance test of the estimator is:

$$MH_{CHISQ} = \frac{\Sigma_j (A_j - \Sigma_j E(A_j))^2}{\Sigma_j Var(A_j)} \qquad (22)$$

where $E(A_j)=N_{Rj}M_{1j}/T_j$ and $Var(A_j)= N_{Rj}N_{Fj}M_{1j}M_{0j}/T^2_j(T_j-1)$.

If X is the dichotomous score on item i, j represents the jth level of total score S, and G indicates the

focal/reference group, the Mantel-Haenszel approach provides the uniformly most powerful test of Ho:

$$H_o: \frac{\frac{P(X_i=1|S=s_j,G=1)}{P(X_i=0|S=s_j,G=1)}}{\frac{P(X_i=1|S=s_j,G=0)}{P(X_i=0|S=s_j,G=0)}} = \alpha = 1 \qquad (23)$$

averaged over K levels. We can also express the relation as $(P_{rj}/Q_{rj})=\alpha\ (P_{fj}/Q_{fj})$ j=1...K. If $\alpha \neq 1$ then the ratio of the probabilities of correct and incorrect responses are not equal between focal and reference groups across K score levels.

A continuity correction improves the calculation of the observed significance levels by approximation of the continuous distribution MH Chi-Square by the discrete distribution Chi-Square (1 df). The formula with continuity correction:

$$MH_{chisq} = \frac{(|\Sigma_j A_j - \Sigma_j E(A_j)| - \frac{1}{2})^2}{\Sigma_j Var(A_j)}. \qquad (24)$$

The MH log odds ratio can be rescaled to place it on the ETS scale of item difficulty by means of $\Delta_{mh}=-2.35\ ln\alpha_{mh}$. This value is the average amount more difficult that a member of the reference group found the studied item than a comparable member of the focal group.

α_{mh} can be shown to be a consistent estimator of the

relation $\alpha=(P_{rj}/Q_{rj})/(P_{fj}/Q_{fj})$ and is the weighted average of the sample cross product ratios in each of the K 2 X 2 tables. The MH procedure is closely related to the Rasch IRT method with ability as a total score or score matched for subgroups. Since $\alpha=(P_{rj}/Q_{rj})/(P_{fj}/Q_{fj})$, $\ln(\alpha)=\ln(P_r/Q_r)-\ln(P_f/Q_f)$. Holland and Thayer (1988) show that if the Rasch model holds, the analysis yields an ability estimate of the parameter Θ for each examinee in each group and a difficulty parameter estimate (b) for the item for each group. These estimates satisfy $\Theta-b_r=\ln(P_r/Q_r)$ and $\Theta-b_f=\ln(P_f/Q_f)$ and with some manipulation $b_f-b_r=\ln(P_r/Q_r)-\ln(P_f/Q_f)=\ln(\alpha)$ and $\alpha=\exp(b_f-b_r)$. When the data follow the Rasch model the Mantel Haenszel null hypothesis in (23) holds under the following conditions:

1) There is no DIF in items other than the studied item. If DIF is present in the studied item then the null hypothesis is false.
2) The matching criterion, M, includes the studied item.
3) The data are random samples of the reference and focal groups.

Proceeding from condition 1, the first step in the recommended procedure is to eliminate biased items based on a preliminary DIF analysis. The matching criterion, M, should be total score on all items left in the refined criterion plus the studied item. If the studied item is left out it is argued that Δ_{mh} will shift by an amount almost independent of the studied item and which depends on overall differences in Θ as estimated by

criterion test between R and F (Zwick, 1990).

The MH and IRT definitions of DIF correspond in the Rasch case because Bayes' sufficiency obtains. We can write the conditional distribution of item scores given θ and S (total score) and the distribution of item scores can be shown to depend on S. The rank orders of θ and S are the same in the Rasch case and it can be said that S contains all the information in θ relevant to the score on the item.

Phillips and Holland (1987) developed a variance formula for $\ln(\alpha_{mh})$ which is valid whenever the numerator and denominator are both large:

$$Var(\ln(\alpha_{mh})) = \frac{1}{2U^2}\Sigma[T_j^{-2}(A_jD_j+\alpha_{mh}B_jC_j)(A_j+D_j+\alpha_{mh}(B_j+C_j))] \quad (25)$$

where $U=\Sigma_j A_j D_j/T_j$.

ETS has developed a classification scheme which is designed to account for both practical and statistical significance of the MH statistics. With some samples, the MH chi-square test may be significant for odds ratios that are insignificant in the practical sense. Therefore there may be a need to control for the within sample variation. The scheme is based on a one sided confidence interval test of the ETS Δ_{mh}, such that it differs significantly from 1, along with other criteria. This ETS significance test evaluates $|SE*\sigma+1|>\Delta_{mh}$ where SE is the one sided confidence interval (.05 level) and $\sigma=((\sqrt{Var})*2.35)$. An item is classified according to the following criteria for low, moderate and high levels of

DIF respectively:

A $|\Delta_{mh}|$ less than 1.0 and/or α_{mh} not significantly different from 1 as per MH chi square test.

B $|\Delta_{mh}|$ greater than 1.0 and α_{mh} significantly different from 1 (MH chi square) and either $|\Delta_{mh}|$ less than 1.5 or not significantly different from 1.0 (ETS significance test).

C $|\Delta_{mh}|$ greater than 1.5 and significantly different from 1.0 (ETS significance test).

F. Mantel-Haenszel DIF Studies

Zwick (1990) evaluated the correspondence of the Mantel-Haenszel and item response theory definitions of DIF. The values of α_{mh} and Δ_{mh} were calculated under a no IRT DIF condition. This approach to detecting DIF evaluates the identical ICC hypothesis which is a comparison of item performance, conditional on overall ability. If X is the dichotomous item score, G is the focal or reference group and ability=θ, then $P(X_i=1|\theta,G=0)=P(X_i=1|\theta,G=1)$ must hold under conditions of no DIF. This is equivalent to the Lord (1980) definition. In general, the MH and identical ICC hypotheses are not the same because observations cannot be stratified on the unobserved θ. Stratification on S or total score is an

approximation of the latent ability continuum.

The relation between the MH (α_{mh}=1, Δ_{mh}=0) and IRT definitions of DIF are explored by Zwick (1990) for a broader class of SMIRFLI (strictly monotone item response functions with local independence) items. It is shown that the MH null hypothesis does not necessarily hold even when a) identical IRF hypothesis holds for every item and b) the matching criterion includes the studied item.

If 1) the distributions $F(\theta|S=s,G=0)$ and $F(\theta|S=s,G=1)$, with S=total score, G=group, are stochastically ordered in the sense that one is monotone increasing on ability at uniformly higher levels than the other and if 2) $P(Xj=1|\theta,S=s,G=g)$ is strictly monotone increasing in θ, then $P(Xj=1|S=s,G=g)$ will take on higher values for the group with the higher ability distribution except when the Rasch model holds. MH will show DIF favoring the group with the higher distribution. When the prior ability distributions are stochastically ordered and the Rasch model does not fit, the Mantel Haenszel procedure will show DIF favoring the higher ability group.

With $b_2=b_1$, Zwick compared MH to the two parameter logistic with three a_2 vs. a_1 conditions including a no difference condition. If the IRF DIF null hypotheses implied that of MH DIF, then Δ_{mh} should equal zero. Δ_{mh} equaled zero only when the two ability distributions were identical or when the discrimination (a) parameters for the two items were identical (when the Rasch model was fit). Departures from values of

zero are greater for larger differences between Group 1 and Group 0 abilities and larger differences in the a_i parameters. Under the Rasch model the MH and IRF DIF null hypotheses held with S, the matching variable, free of biased items and including the studied item. S contained all the information necessary for estimating examinee abilities, functioning as a sufficient statistic for θ.

In a related investigation in this study it was demonstrated that if the studied item is excluded from the matching criterion MH will show DIF favoring the group with the higher distribution. In summary, it appears that the MH null hypothesis will not hold, in general, if the ability distributions are stochastically ordered. It will only hold, under these conditions, under the Rasch model. An additional finding is that lower reliabilities of total score will result in larger discrepancies between the IRF and MH definitions of DIF.

Linacre and Wright (1987) criticize the use of the MH procedure, maintaining that it is statistically inferior to the Rasch approach. They argue analytically that the value of α_{mh} is affected by the distribution of abilities, selection of interval boundaries and absolute sizes of focal and reference groups. The number of ability levels is "arbitrary" and therefore the statistic does not have meaning beyond the particular matching scheme used. As previously discussed, under the Rasch model $\theta-b_r=\ln(P_r/Q_r)$, $\theta-b_f=\ln(P_f/Q_f)$ and b_f-

$b_r=\ln(P_r/Q_r)\ln(P_f/Q_f)=\ln(\alpha)$. The difference between the estimates b_f and b_r is independent of the ability distribution of both groups since θ drops out in the above relation. Therefore, odds ratio estimates are independent of ability levels of the two groups under the Rasch model and the comparison of b_f and b_r under the Rasch model does not require any matching of ability levels.

They also argue that MH_{CHISQ} is the estimate of the statistical significance of the difference between the performance levels of the reference and sample groups. In segmentation and matching schemes the distribution of abilities, selection of boundaries and absolute sizes of reference and focus groups must also affect the size of MH_{CHISQ}. It can be shown that the significance statistic is estimating $CHISQ=(\Sigma(P_{rj}-P_{fj})N_{rj}N_{fj}/T_j)^2/\Sigma(Var(A_j))$. The MH significance statistic is therefore obtained by averaging over different "arbitrarily" defined ability levels the difference between the groups in the probability of a correct response to an item. Thus, the nature of the matching schemes used can affect the stability of the MH_{CHISQ}.

Shultz (1989) reports that MH is less sensitive to the detection of DIF than the Rasch procedure with small groups. Modelled DIF successfully detected by both the MH and Rasch procedures with n=1000 was not recovered by MH when n=100. However, this may be a result of the method by which the data were simulated, which was not specified. Their results also

indicated that the Rasch model is more reliable than MH. Several MH variants were tested in which a) the studied item was excluded or included in the criterion used for matching b) the matching was fat or fine. When groups actually differed in ability, with Θ as total score, only the MH procedure which included the studied item and used all levels of the total score for matching produced no "net DIF" (confusion with real group differences) over items. In this study the correlation of MH and Rasch prodedures for detecting DIF was at its theoretical maximum (.99) for equal achievement contrasts, that is for matched groups. For unequal contrasts the correlation was .81, suggesting a dependency in the MH statistics related to item X achievement (ability) interactions It is suggested that these interactions are readily detected by the Rasch item fit statistics.

Hambleton and Rogers (1989) demonstrate that the Mantel-Haenszel statistic can be considered analogous to comparing two ICCs based on a one parameter model. It is not designed to detect non-uniform bias, that is in which the ICCs cross. In this situation an item is biased against one group at low Θ and against the other at higher Θ. In the uniform bias situation the ICC for one group is higher than or equal to that for the other group across the range of Θ. Widely differing ability distributions will affect IRT bias statistics by affecting parameter estimation. The effect of discrepant score distributions on the Mantel Haenszel

statistic may be more difficult to predict. With two Anglo-American and two Native-American samples of 1000 subjects each of real data independent bias analyses were carried out to evaluate for consistency. A matched group analysis was performed for the Mantel-Haenszel statistic by selecting a Native American sample with a score distribution that closely matched the Anglo-American sample. The three parameter model was fit with small residuals to the data for the IRT DIF analyses.

Both methods displayed considerable instability across samples. For the IRT based analysis 73% of items were detected as biased in both analyses. The corresponding frequency was 80% for the MH procedure. The authors consider this "disturbing" in view of the fact that in most situations one would not have the luxury of a cross-validation sample. The two methods were then compared with items which were consistently identified as biased across samples. For the unmatched analyses, of the 14 items consistently identified by the area method and the 9 items consistently identified by MH, 7 items were in common. The items consistently flagged by one method but not the other were then examined. For 4 of these items the ICCs crossed markedly. It is not surprising that MH did not detect these. The Mantel-Haenszel statistic was found to be affected by the numbers of examinees in certain ability ranges for some items regardless of the ICCs marked or small area differences. A modified area statistic over a restricted

interval was designed to identify fewer biased items with crossing ICCs. This resulted in a closer correspondence between the MH and ICC methods under these circumstances. The modified area statistic was more closely related to the MH results.

The matched sample analysis results were similar for the MH procedure in that all the previously detected items were identified. In addition, four others that had been flagged in the unmatched analysis were detected as was one previously undetected item. Thus, both methods are somewhat unreliable in identifying biased items. Studies of item bias are influenced to a large degree by the sampling error of item bias statistics. There was substantial agreement in the identification of uniformly biased items but the MH procedure did not identify non-uniformly biased items. Restricting the ability interval to the range where most of the focal group members score caused the MH method to show more congruence with the IRT results. The distribution of test scores had little impact on the MH results and MH results remained stable with smaller sample sizes. The authors conclude that the MH method can be safely substituted for the ICC method if safeguards are put in place to detect non-uniformly biased items.

Donaghue et al (1989) conducted a simulation experiment in which several of the factors expected to influence the magnitude of Δ_{mh}, a function of α, were varied under a three

parameter logistic IRT model. These were then analyzed in an ANOVA design. The various conditions violated the assumptions of Holland and Thayer (1988) in several ways. They delineated the following factors that may impact on Δ_{mh}, through α_{mh} when the assumptions are not met: 1) Reliability/Dimensionality of M, the matching criterion items. 2) Distribution of item characteristics in M. 3) Amount of IRT bias in items comprising M. 4) Distribution of M in Reference and Focal groups. 5) Grouping of M. 6) Deletion of items based on preliminary analyses. 7) Sample sizes of R and F. 8) Characteristics of studied item for R and F. They varied 6 factors which included: I. Between Data Sets A-a parameter-shared between studied and criterion items (3 levels), N-number of biased criterion items in the matching variable in M (4 levels). II. Within Data Sets B-actual bias in studied item (2 levels), D-Difficulty of studied item in the reference group (3 levels). III. Mode of Analysis I-inclusion of studied item (2 levels), J-number of items in matching criterion. Thus there were 576 conditions specified. Number of subjects was 2000 in the Reference group and 500 in the focal group. θ for R and F were sampled from N(0,1) and N(-1,1) respectively.

There were large main effects for I, B, and D, in respective order. The fact that the inclusion/exclusion of the studied item had a stronger effect than the actual bias in the item was interpreted as convincing evidence that including

the studied item is the correct procedure in MH analysis. Since B represented the actual amount of DIF in the studied item it is appropriate that was significant. The significance of the D main effect should be considered in relation to the locations of the reference and focal group abilities, although this goes beyond the interpretation of the authors. Increasing D (-.5, 0, .5) moves the item away from the combined distribution of reference and focal group abilities. It is suspected that ability discrepancy plays a role in the significance of this factor, but the authors did not vary the ability discrepancy between the groups.

Various interactions also had important effects, among these was AD. In this context it was noted that increasing a overestimated the Δ_{mh} values. Increasing D underestimated them which was attributed to the decision to model a non-zero guessing parameter $c_r=c_f=.20$. Increasing D decreased Δ_{mh} both as a main effect and in its interactions with B and A. Although A had a small main effect it showed strong interactions with B and D. N evidenced small main and interaction effects, the effect of bias in the matching variable was most acute with increasing A. When J was small (4 or 9 items) Δ_{mh} showed greater sensitivity to D. Increasing J tended to reduce the impact of the other factors.

It was concluded that 1) The studied item should always be included in MH analysis because MHD and its standard error will be significantly more accurate than if the studied item

is not included. 2) If $c_r=c_f$ MH analysis will be less likely to detect a given amount of bias in a hard item than an easy one. 3) There is only a small influence of IRT bias in the matching variable.

Spray (1989) conducted a series of simulation trials using actual ACT test item responses. Ability levels based on total score were replaced with a latent ability measure estimated from the data which was successfully modelled under the three parameter logistic model. Several problems are discussed with regard to conditional item bias or DIF statistics. First, ability estimates based on a total score contain some measurement error which can be explained as a function of the reliability of the test. Thus subjects with the same "true" ability can have different total scores or be placed in different score groups.

Secondly, subject data is obtained from available sample sizes for reference and focal groups. These may be widely discrepant and may not provide stable referents, especially for a considerably smaller demographic group. Ideally we might consider matched ability groups but these will often require larger samples to properly represent certain groups.

Thirdly, item scores are aggregated to arrive at a total score. A small, constant bias in many of the items in the matching criteria may cause the overall test to evidence considerable bias. The MH procedure cannot detect this type of constant bias as it is a significance test for individual

items. Under the Rasch model this can be evaluated by the between infit statistic as outlined in Smith (1988). This statistic tests the fit of a particular items by examinees response matrix to the dichotomous Rasch model. It aggregates the standardized residuals between the observed and expected scores predicted by the model for L items and across M inhabited score groups. This statistic is approximately chi square with (L-1)(M-1) degrees of freedom. It is possible for no individual items to show bias as per the MH statistical test and yet for the test to be considerably biased. This would be the case if small or moderate standardized residuals were aggregated over many items. The between fit statistic shows substantial power and sensitivity in the detection of systematic measurement disturbances such as DIF.

Spray (1989) also found that differences in the distribution of frequencies along the range of abilities or, "where the focal group ability is distributed", affected the performance of the MH statistic. It was also found that the MH estimator overestimated the true MH value for smaller ($n<250$) samples and underestimated it for large samples (≈ 2000). MH was found most effective for moderate sample sizes and test lengths. The relationship between variability of the odds ratio estimator and sample size was evaluated through relative efficiency: $R.E.=n_2\sigma_2^2/n_1\sigma_1^2$. Decreases in sample size which produced values close to 1 indicated no special or unusual diminution of accuracy other than what was

expected by the sample size reduction. However, there were severe departures from unity for MH when sample size was reduced from 250 to 100 and 100 to 50 subjects. This indicated unstable and accelerating losses of relative efficiency as sample sizes were reduced. This suggests that MH is effective for moderate sample sizes but would not be recommended for small samples.

Camilli and Smith (1990) demonstrated in a simulation study that the MH chi square significance levels are extremely robust. They compared MH results with those obtained by a randomization test and a jackknife test and found a high degree of correspondence. In an exact significance test all possible permutations of each 2 X 2 table within the 2 X 2 X K layout were considered. The marginal frequencies were fixed and A_j set free to vary. The permutations of the 2 X 2 X K layout are labeled $A_1', A_2' \ldots A_N'$. The exact significance level is the proportion of K-tuples with $\Sigma^k A_j' > \Sigma^k A_j$. The randomization test procedure utilized in this study is an approximation to the exact test in which the K-tuples are sampled with a random hypergeometric number generator. The randomization procedure performed very similarly to the MH chi-square statistic in the context studied in this paper. Another important finding from the randomization phase of this study is that DIF is found more easily in highly discriminating items than less discriminating items once ability is accounted for. Another bias index can be formed by

dividing $\ln(\alpha_{mh})$ by its standard error as calculated from an approximation formula such as from Phillips and Holland (1987). The standard error can also be obtained by jackknifing as in this study. Jackknifing can be particularly useful when there are unanticipated, influential sources of variation in the data. Δ_{mh} frequently exhibits this type of variation with irregular patterns of positive and negative values across ability categories taken as total score. In this study, ability subgroups were individually deleted and Δ^* calculated and transformed into pseudo values. For each subgroup: $\Delta^*=K\cdot\Delta_{all}-(k-1)\cdot\Delta_{(j)}$. The mean and standard error of these pseudo values were used to calculate the jackknifed statistic, which is approximately distributed as a t statistic. The t statistic was then converted into a percentile and compared to the nominal MH statistic. The jackknifed and nominal levels were in close correspondence further indicating the robustness of the MH chi-square.

Harvey (1990) examined the stability of the MH procedure across samples differing in ability. In this study real data from the June and December administrations of the College Board Admissions Testing Program (ATP) achievement test were analyzed. Examinees at the two administrations differ in terms of academic preparation and therefore the proportion of examinees at various ability levels differs between administrations. Several items were found to have very different MH DIF statistics and classifications between the

two administrations which seemed to be the result of these changing proportions. Differences in the DIF statistics appear to be due to different relative abilities of the groups. Although the MH statistic is weighted to account for these differences, MH analyses could be discrepant if there is more DIF in one portion of the ICC than the other with changing ability distributions. Data were not presented regarding ICCs in this study. This seems to be the most reasonable explanation for the discrepant MH results.

Swaminathan & Rogers (1991) proposed the use of a logistic regression based procedure for detecting DIF. It was suggested that the procedure is more general and flexible than the MH procedure. It is capable of detecting nonuniform DIF but is 3-4 times as expensive as the MH procedure. The probability of a correct response for person i in group j given θ is predicted. θ is taken as total score, hence logistic regression evaluates observed score invariance (OCI). The probability is predicted by

$$P(u_{ij}=1|\theta_{ij}) = \frac{e^{(\beta_{0j}+\beta_{1j}\theta_j)}}{[1+e^{(\beta_{0j}+\beta_{1j}\theta_j)}]}, \tag{26}$$

which specifies separate equations for the two groups of interest. β_{0j} is the intercept parameter and β_{1j} is the slope parameter for group j. θ_{ij} is the ability of individual i in group j. No DIF is present if the logistic regression curves for the two groups are equivalent. This will be the case if

$\beta_{01}=\beta_{02}$ and $\beta_{11}=\beta_{12}$. If $\beta_{11}=\beta_{12}$ and $\beta_{01}\neq\beta_{02}$, the regression curves are parallel and uniform DIF is present. If $\beta_{01}=\beta_{02}$ and $\beta_{11}\neq\beta_{12}$ the curves intesect and nonuniform DIF is present. Alternative equivalent formulations of the model can be devised based on various coding schemes.

The authors conducted a simulation study comparing the logistic regression procedure with the MH procedure. They concluded that the logistic regression procedure is as powerful as the MH procedure in detecting uniform DIF and more powerful in detecting nonuniform DIF.

G. Summary and Purpose

Several issues appear to have been satisfactorily resolved in the MH DIF literature and are not the subject of this work. It appears that the impact of sample size of the R and F groups and the total number of test items on the accuracy and variability of the MH odds ratio are understood, (Schultz 1989; Spray 1989). It is also clear that the MH procedure is not designed to detect the presence of non-uniform bias, that is cases in which the R and F ICCs cross (Holland & Thayer, 1988; Swaminathan & Rogers, 1989). A clear consensus appears to have emerged that the studied item should be included in the matching criterion and that the presence of a small number of biased items in M will not significantly degrade the MH statistic (Holland & Thayer, 1988; Zwick, 1990; Donaghue et al, 1991).

The results of the previous studies indicated that some

important issues regarding the influences on the MH procedure in DIF analysis required resolution.

i) Research Phase I

When response data follow the Rasch Model, the studied item is included, and the other items are unbiased, the MH odds ratio is independent of ability distributions of reference and focal groups. Zwick (1990) has shown that the MH and IRT hypotheses will not be necessarily be identical for SMIRFLI items (strictly monotone item response functions with local independence). While Rasch items are SMIRFLI, a broader SMIRFLI class includes items following two and three parameter IRT models. The validity of application of the MH odds ratio estimator and chi-square significance test in DIF analyses in data sets following two and three parameter models is called into question as the MH and IRT DIF hypotheses are not identical. The effects and interactions of the discrimination, difficulty, pseudo-guessing and group ability distribution parameters should influence the accuracy of the MH statistics according to Zwick (1990). This in turn may influence decisions about DIF in an item. Many data sets encountered follow two and three parameter IRT models. Frequently the reference group has a higher distribution of abilities than focal groups. Research phase I was designed to examine the effects of item parameterizations and discrepant ability distributions of the focal and reference groups on the MH odds ratio estimator and chi-square significance test.

Since the Educational Testing Service has proposed the use of its DIF classification scheme it was decided that these would be computed and compared to the nominal MH chi-square levels in the above analyses. The rationale was to evaluate whether the stability of the classification scheme is affected under the conditions varied in the study.

ii) Research Phase II

The identical nature of the MH and IRT definitions of DIF, in terms of the odds ratio, α, does not necessarily hold for the MH chi-square significance test. Empirical work evaluating whether discrepant ability distributions affect the accuracy of the MH procedure through effects on the power of the significance test has been missing. In a statistical sense, the precision of an estimator (α_{mh}) does not necessarily speak to the power to detect the estimator as significant.

The effects of discrepant ability distributions and difficulty parameters of the studied item on the performance of MH chi square in the Rasch case have never been fully examined. It has been demonstrated that the MH and IRT definitions of bias in terms of the MH odds ratio, α, are the same in the Rasch case. However, a more complete understanding of the relationship between the two approaches in the Rasch case is required as it is insufficient to demonstrate that α behaves as expected. The power of the chi-square test to detect α as significantly different from 1 must be examined in this case. Thus, research phase two is

designed to examine the effect of discrepant ability distributions and difficulties of the items in the matching criteria on the MH chi-square statistic.

iii) Research Phase III

Another area in which investigation has been required is the performance of the empirical and estimated (Phillips & Holland 1987; Holland & Thayer, 1988) variances of the MH odds ratio. Little is known regarding the effects of parameterizations and ability distributions on the variance of the MH odds ratio in the DIF application. Donaghue et al (1991) evaluated MH-RATIO (Var_{emp}/Var_{est}) in the ANOVA framework of their study. They found that the estimated variance (Var_{est}) overestimated the empirical variance (Var_{emp}) by approximatly 4% overall. The b parameter factor (D) of the studied item was a significant main effect. This suggests a need to evaluate the variances while varying the item parameter factors and ability discrepancy factor (not studied in Donaghue et al). Therefore the variances derived from research phase I provide important information for carrying out these analyses.

iv) Research Phase IV

Several studies also referred to the problem of the unreliability of test items in relation to the MH procedure (Spray, 1989; Zwick, 1990). The MH procedure utilizes total score as the measure of ability. The extent to which this is reasonable in DIF analysis is an important issue which has

not been evaluated comprehensively. Therefore, the reliability of total score as a measure of θ required assessment. To the degree that persons' total scores and levels of θ are discrepant the MH statistic and our decisions about DIF are likely to be influenced. The nonlinear regression of θ on total scores can be evaluated as an expression of the reliability of total score as a proxy for θ. Total score can be evaluated in terms of the variance accounted for in θ and that quantity expressed as an eta^2 statistic. Eta^2, like reliability, should approach unity for an infinitely long test. Therefore increasing the number of items increases eta^2. The distribution of parameters in items on the test other than the studied item, that is in M, will also directly affect eta^2. This issue has yet to be examined and is considered an important area for research. The impact of item parameterizations if the items in M is hypothesized to affect the MH statistic through their effects on reliability in the sense of eta^2. The effects of these factors are examined in phase 4.

Since the MH odds ratio is a measure that can quantify DIF it has an important advantage over earlier chi-square methods. In addition, the MH chi-square test statistic is sensitive to specific inequalities in the odds ratio relationship , not just a general association. The deviation assessed in DIF studies is of a type that the MH statistic is sensitive to.

This study is conducted to assess the general issues of the influences of 1) ICC and ability parameters, 2) the reliability of total score as a measure of θ, on the MH odds ratio, MH chi-square test, estimated and empirical variances and ETS DIF classification statistics. This work also specifically evaluates the equivalence of the MH and IRT DIF definitions of DIF in the Rasch case. This is done in a systematic, controlled way in simulated data.

III. Method

Simulated data sets were generated from logistic item response functions by use of RANGEN (Millsap, 1990) in which a, b and c parameters for each item are input for one, two and three parameter models. Ability distribution, test length and subject sample size and a seed value from which data generation begins are also specified. Abilities are defined through a normal prior distribution with variable mean and standard deviation. A person by items matrix of item scores is generated and the program produces theta values for each subject and scales them.

For each simulated set of conditions the program regresses theta values on total score and calculates the variance of theta conditional on total test score for each reference and focal group data set. For a user defined number of replications, the program calculates and outputs the mean, variance and additional descriptive statistics of the eta^2 statistic for each group. The program also outputs these statistics for α, ln α, variance of α (Phillips & Holland, 1987; Holland & Thayer 1988), MH chi-square, and ETS Δ. An additional subroutine calculated the number of replications for which the studied item was placed in the ETS A, B and C DIF classifications. For all phases of the study the matching criterion was total score including the studied item. There was no bias in any item except for possibly the studied item. Additional information about RANGEN is given in Appendix A.

Phase I.

The first phase of this study evaluated the effects of item parameterization of the studied item and ability distribution discrepancies on the MH odds ratio and chi-square significance test. The set of studied variables for the studied item was determined through a 3 x 3 x 2 x 2 design. The studied factors were: 1) A, the discrimination parameter, (3 levels 0.5, 1.0, 1.5), 2) B the difficulty parameter, (3 levels 0.0, 0.3, 0.5), 3) C the pseudo-guessing parameter, (2 levels 0.0, 0.20), 4) Θ_{dis}, the group discrepancy in mean ability level (2 levels 0.0, -1.0). The no-discrepancy in ability distributions condition was included to evaluate the impact of differences in ICC parameters between groups independent of ability distributions on the accuracy of the MH procedure.

For each combination of variables for the studied item there were 200 replications for both biased and an unbiased condition. For the unbiased condition the item response functions were identical for the reference and focal groups. The parameter values were equal to the cell factor values. For the biased condition the focal group IRF was defined by the cell factor values and the reference group IRF parameters were a=1, b=0, c=.20. This defined the various discrepancies in ICC parameters between focal and reference groups which were evaluated. Moderate sample size and test length were used of n=500 examinees for each group and k=20 items. There

was no bias in matching criteria items. The ICC parameterizations of matching criteria items covered a range of a parameter values (.5, 1.0, 1.5) representing low, moderate and high discrimination, b parameter values (-2.25 to 2.25, .25 intervals) covering the full range of difficulty values and c parameter values (0, .10, .20) representing levels commonly encountered in practice.

Each combination of studied variables in the design produced two analyses, a type I error analysis for the no bias condition and a type II or power analysis for the biased condition. The sampling distribution of the α, MH chi-square and ETS transformed statistics was evaluated in terms of type I error by examining the frequency of rejection of the null hypothesis of no bias. In the type II error or power analysis frequency of failure to reject the null hypothesis when bias was modelled was examined. The critical values for these analyses were determined by the MH chi-square test of association p-value (.05) and the ETS transformed MH odds ratio statistic $\Delta_{mh}=-2.35\ln\alpha_{mh}$. The ETS classification scheme cutoff values are $\Delta_{mh}\leq 1.0$=low DIF, $1.0\leq\Delta_{mh}\leq 1.5$=moderate DIF, $\Delta_{mh}\geq 1.5$=high DIF. The use of the these values permits an evaluation of the decisions ETS would make regarding the presence of DIF under the studied conditions, (Harvey, 1990).

In both the type I and type II analyses the variance of the sampling distributions of the MH statistics is obtained and is of considerable interest with regard to the error rates

and ETS DIF classifications. The variance of $\ln\alpha$ is calculated both empirically and through the Holland & Thayer (1988) formula, herein referred to as $Var_{ht}\ln(\alpha)$. An assessment is then made regarding of the extent to which the estimated variance formula approximates the actual variance of $\ln\alpha$ under varying conditions.

ANOVA analyses were utilized to identify the significant factors affecting α and MH_{CHISQ}, in the type I and type II error rate cases, among the ICC parameter values and the Θ_{dis} factor manipulated in the individual item simulations. The ANOVAs were performed with α or MH_{CHISQ} (type I or type II) error rate as the dependent variable and the A, B and C parameter factors and the Θ_{dis} ability factor as the independent variables. This enabled a summary statistical analysis from the fixed paradigm of no DIF in any of the studied items and in the context of the variable DIF modelled in the study.

Phase II.

In the second phase of the study the Rasch case is further explored. This is done to study the behavior of MH chi square under various conditions in which the amount of bias in the studied item is constant but the characteristics of M, the matching criterion and the reference and focal group ability distributions is varied. In this phase the reference group ICC is set at (a, b, c: 1, 0, 0) and the focal group 1, .3, 0. Thus, a difference of .3 in the difficulty parameter of the two ICCs from which data were generated was sought in the MH

statistics for each case. The mean value of the Θ distributions for the focal and reference groups and the difficulties of the items in the matching criteria were increasingly discrepant in each case. The items in M were separated by intervals of .25 on the difficulty dimension. The studied values for mean Θ for each group and the spread of difficulties (b parameters) for items in the matching criterion, M, are as follows:

Case	Θ_r	Θ_f	Difficulties of items in M by .25
1	0.0	-1.0	-2.25 to 2.25
2	0.0	-1.0	-1.25 to 3.25
3	0.0	-1.0	-1.00 to 3.50
4	0.0	-1.0	-0.75 to 3.75
5	-0.5	-1.5	-0.75 to 3.75
6	-0.5	-1.5	0.00 to 3.75[1]

The complete set of eta^2 and MH statistics generated by RANGEN and specified previously was obtained for these runs.

Phase III.

In the third phase of this study the influences of the phase I factors on $Var_{ht}\ln(\alpha)$ and the empirical variance of $\ln(\alpha)$ (referred to as $Var_{emp}\ln(\alpha)$) of the studied items are examined through a set of ANOVAs. Variance data were utilized from the first phase of this study and both the type I and type II cases were examined. In the first two cases, ANOVA analyses were carried out under the type I error rate

[1]Two items each for b=3.50 and b=3.75

condition of no DIF with $Var_{ht}\ln(\alpha)$ and $Var_{emp}\ln(\alpha)$ as the dependent variables, respectively. The independent variable factors are as in the first phase of the study, A, B and C parameter factors and the Θ_{dis} factor. This was designed to examine the influences of the factors on the absolute magnitudes of the variances in the type I case. In the next two cases, the corresponding analysis was carried out for the type II cases of the varied modelled DIF conditions. In the last two cases the discrepancy between the variances was examined through calculating

$$DISC_{var} = [\frac{Var_{ht}\ln\alpha}{Var_{emp}\ln\alpha} - 1] .$$

This quantity permits an interpretation of $Var_{ht}\ln(\alpha)$ as a percentage of the empirical variance. Using this quantity as the dependent variable ANOVAs were carried out for the type I and type II error rate cases with the independent variable factors described for this phase.

Phase IV.

In the fourth and final phase of this study the reliability of total score as an indicator of theta was examined as it impacted on the MH statistics. This measure is closely related to the information function in an IRT model. In utilizing the RANGEN program we generate known Θ or person parameter values and have S or total score information available. Therefore we can directly examine the relationship

$Var(\theta|S)$. This is of interest because it addresses the question of the adequacy of S as a proxy for θ.

A RANGEN program subroutine regresses θ on total scores and generates an eta^2 value which expresses the amount variance of the scaled θ value explained by total score. Eta^2 is obtained:

$$\eta^2 = 1 - [\frac{Var(\theta|S)}{Var(\theta)}] .$$

This quantity depends on $Var(\theta)$ given, knowing S. When $Var(\theta|S)=0$ the Mantel-Haenszel should correspond perfectly to the IRT DIF formulation.

The parameter values of the items in M influence eta^2 through their effects on the information functions. The level of steepness of a parameters, spread of item b parameters and levels of c parameter values influence eta^2. Steep a parameters should produce more information at certain θ values depending on the value of b. These b values can be spread across the θ range providing high overall levels of information. The presence of c parameters will tend to degrade the amount of information at least at the low end.

This phase focused on the level of eta^2 and the accuracy of the MH statistics. Eta^2 was manipulated indirectly by varying the parameter values of the items in the matching criteria, M. These are known to influence reliability of items and total score. The research examined some of the type I and type II error analyses possible subsequent to manipulation of eta^2

values through varying item parameters in M.

In limited pilot testing eta^2 was affected by individually manipulating the discrimination, difficulty, pseudo guessing, test length and ability factors (see Table 1).

TABLE 1

Eta Manipulations

A	B	C	Mean Θ	# items	Eta	Eta^2
1.0	-2.50 to 2.25	0	-0.5	20	.859	.738
1.5	-2.50 to 2.25	0	-0.5	20	.910	.828
2.0	-2.50 to 2.25	0	-0.5	20	.934	.872
2.0	-2.50 to 2.25	.20	-0.5	20	.800	.640
1.0	-1.35 to 1.50	0	0.0	20	.871	.760
1.0	-1.35 to 1.50	0	-0.5	20	.867	.752
1.0	-1.00 to .90	0	0.0	20	.880	.775
1.0	-1.00 to .90	0	-0.5	20	.863	.745
1.0	0.0 to 2.80	0	0.0	20	.804	.646
1.0	0.0 to 2.80	0	-0.5	20	.716	.513
1.0	-2.50 to 2.50	0	0.0	21	.873	.763
1.0	-2.50 to 2.50	0	0.0	16	.813	.662
1.0	-2.50 to 2.50	0	0.0	11	.743	.552

It is also very likely that there are important interactions

among these factors. For example, the effect of discrepant ability distributions on eta^2 appears to depend on the spread of the item difficulty parameters in the set of items.

If total score is a good proxy for Θ then the MH statistics should correctly identify actual bias and no bias conditions simulated in item response functions. Findings regarding the level of eta^2 required for accurate decisions in Mantel Haenszel DIF analysis were of particular interest. In order to affect eta^2, a and c parameter values for items in the matching criteria were manipulated along with the Θ discrepancy between the reference and focal groups. Item b parameter values covered the range -1.80 to 1.80 by intervals of .20 in all cases. In the interest of simplicity test length was set at 20 items and number of examinees were set at 500 for both the focal and reference groups. There were 5 levels of a (.25, .40, .75, 1.25, 1.75), 3 levels of c (0, .10, .20) and 2 levels of Θ_{dis} (0, -1.0). Thus there were 30 combinations of variables from which the item X examinee data matrix was generated. Each combination produced a type I and type II error analysis. Thus this phase of the study consisted of a total of 60 runs. Again, there were 200 replications for each set of studied variables. In the no DIF cases the studied item had parameter values of (1, 0, 0), for A and C ICC parameter values equal for all items. Θ (ability) is normally distributed, St.Dv.=1.0. B values dispersed at equal intervals.the focal group, for a type I error rate

condition. For the DIF cases the focal group studied item parameter values were (1, .3, 0) and type II error rates were evaluated. The level of eta for the reference and focal groups was indirectly manipulated through the item parameters in M and the θ_{dis} factor. Therefore, of primary concern is the level of eta of each cell of the design for each group, taken as the mean across the focal and reference groups. Two multiple regression analyses were then performed with type I and II error rates as the dependent variables. This was done to evaluate the statistical significance of the independent variables among θ_{dis} (factor), eta (continuous covariate) and θ_{dis} by eta (interaction).

IV. Results

Phase I.

The results of the studied item parameterization analyses are presented in a series of tables. Tables 2 and 3 reflect MH chi-square and ETS DIF classifications for type I and type II analyses, respectively. In Tables 4 and 5 the Mantel Haenszel alpha (α_{mh}) and ETS delta (Δ_{mh}) values are presented for type I and type II analyses, respectively. In Tables 6 and 7 the empirical variance, $Var_{emp}ln(\alpha)$, the estimated variance, $Var_{ht}ln(\alpha)$, from Holland and Thayer (1988) and the percentage discrepancy between them are presented for the type I and type II analyses.

ANOVAs are calculated using θ_{dis} and item parameter values as independent variable factors and the MH chi-square type I and type II error rates and the odds ratio estimator α as the dependent variable. These are utilized to guide and summarize the findings.

Cases are specified by the underlying ICC parameters and θ_{dis} values for the reference and focal groups from which RANGEN generated the data for the MH analyses. The ICCs defined by the parameter values were plotted using MATHCAD (Addison & Wesley, 1988) and are presented in Figures 1 through 20. The areas between these curves give an indication of the amount, form and type of DIF modelled. Cases are referred to by their row and column positions in Tables 2, 3, 4, 5, 6 and 7. For example, case MH24 is in row 2, column 4,

TABLE 2

MH Type I Error Rate[1]
ETS A[2], B[3] and C[4] DIF Classifications
No DIF Modelled A Conditions

	θ_{dis}=0.0						θ_{dis}=-1.0					
	c=.00			c=.20			c=.00			c=.20		
	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50
a=0.5	.055[1] A-189[2] B- 11[3] C- 0[4]	.035 A-193 B- 7 C- 0	.03 A-194 B- 6 C- 0	.03 A-194 B- 6 C- 0	.045 A-191 B- 9 C- 0	.05 A-190 B- 10 C- 0	.03 A-194 B- 6 C- 0	.065 A-187 B- 12 C- 1	.075 A-185 B- 15 C- 0	.125 A-175 B- 25 C- 0	.095 A-181 B- 19 C- 0	.125 A-175 B- 25 C- 0
a=1.0	.045 A-191 B- 9 C- 0	.04 A-192 B- 8 C- 0	.025 A-195 B- 5 C- 0	.03 A-194 B- 6 C- 0	.055 A-189 B- 11 C- 0	.045 A-191 B- 9 C- 0	.095 A-181 B- 19 C- 1	.055 A-181 B- 17 C- 2	.04 A-189 B- 11 C- 0	.05 A-192 B- 8 C- 0	.05 A-190 B- 10 C- 0	.08 A-184 B- 16 C- 0
a=1.5	.055 A-189 B- 11 C- 0	.04 A-192 B- 8 C- 0	.03 A-194 B- 6 C- 0	.03 A-194 B- 6 C- 0	.03 A-194 B- 6 C- 0	.04 A-192 B- 8 C- 0	.19 A-163 B- 32 C- 5	.14 A-173 B- 21 C- 6	.16 A-169 B- 19 C- 12	.06 A-189 B- 11 C- 0	.035 A-193 B- 7 C- 0	.065 A-187 B- 13 C- 0

Parameter values above are for both Focal and Reference groups.

TABLE 3

MH Chi Square Type II Error Rate[1]
ETS A[2], B[3] and C[4] DIF Classifications
Modelled DIF B Conditions

	θ_{dis}=0.0						θ_{dis}=-1.0					
	c=.00			c=.20			c=.00			c=.20		
	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50
a=0.5	.08[1] A- 16[2] B-153[3] C- 31[4]	.00 A- 0 B- 28 C-172	.00 A- 0 B- 3 C-197	.96 A-192 B- 8 C- 0	.71 A-142 B- 58 C- 0	.365 A-173 B- 23 C- 4	.395 A- 79 B-117 C- 4	.025 A- 5 B-115 C- 80	.005 A- 1 B- 29 C-170	.68 A-136 B- 60 C- 4	.93 A-186 B- 14 C- 0	.94 A-188 B- 12 C- 0
a=1.0	.085 A- 17 B-117 C- 66	.00 A- 0 B- 1 C-199	.00 A- 0 B- 0 C-200	N/A	.385 A- 77 B-114 C- 9	.04 A- 8 B-140 C- 52	.015 A- 3 B- 57 C-140	.00 A- 0 B- 0 C-200	.055 A- 0 B- 0 C-200	N/A	.63 A-126 B- 73 C- 1	.21 A- 41 B-151 C- 8
a=1.5	.02 A- 4 B-114 C- 82	.00 A- 0 B- 0 C-200	.00 A- 0 B- 0 C-200	.96 A-192 B- 8 C- 0	.24 A- 48 B-133 C- 19	.00 A- 0 B- 78 C-122	.00 A- 0 B- 2 C-198	.00 A- 0 B- 0 C-200	.00 A- 0 B- 0 C-200	.91 A-182 B- 18 C- 0	.29 A- 58 B-121 C- 21	.065 A- 13 B-111 C- 76

ICC parameter values above are for Focal group.
Reference group ICC (1, 0, .2)

TABLE 4

α_{mh}^{1}, Δ_{mh}^{2}

No DIF Modelled A Conditions

	θ_{dis}=0.0						θ_{dis}=-1.0					
	c=.00			c=.20			c=.00			c=.20		
	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50
a=0.5	1.02[1] -0.01[2]	1.02 -0.02	1.02 -0.04	1.00 .03	1.00 .02	1.01 .00	.96 .13	.93 .21	.96 .14	.90 .28	.89 -.30	.89 -.12
a=1.0	1.04 -0.05	1.01 -0.01	1.01 .01	1.01 -0.01	1.02 -0.01	1.02 -0.02	1.16 -0.22	1.12 -0.20	1.12 - .23	1.01 .00	.97 .12	.95 .16
a=1.5	1.03 -0.03	1.01 .00	1.01 .03	1.00 .04	1.02 -0.01	1.00 .03	1.29 -0.54	1.30 -0.57	1.30 -0.55	1.06 -0.09	1.01 .01	.95 .17

Parameter values above are for both Focal and Reference groups.

TABLE 5

α_{mh}[1], Δ_{mh}[2]
DIF Modelled B Conditions

	θ_{dis}=0.0						θ_{dis}=-1.0					
	c=.00			c=.20			c=.00			c=.20		
	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50
a=0.5	1.66[1] -1.17[2]	2.22 -1.85	2.64 2.25	1.01 -0.01	1.26 -0.52	1.42 -0.80	1.44 -0.83	1.88 -1.45	2.26 -1.88	.79 .58	.95 .14	1.06 -0.12
a=1.0	1.80 -1.35	2.83 -2.41	3.92 -3.18	N/A	1.45 -0.86	1.75 -1.30	2.11 -1.71	3.43 -2.86	4.75 -3.62	N/A	1.31 -0.61	1.53 -0.98
a=1.5	1.85 -1.42	3.31 -2.78	5.00 -3.74	1.00 .03	1.55 -1.00	1.98 -1.59	2.79 -2.38	5.04 -3.76	7.80 -4.77	1.11 -0.22	1.58 -1.01	1.81 -1.37

ICC parameter values above are for Focal group.
Reference group ICC (1, 0, .2)

TABLE 6

$Var_{ht}\ln(\alpha)$[1]
$Var_{emp}\ln(\alpha)$[2]
$(Var_{ht}\ln(\alpha)/Var_{emp}\ln(\alpha))-1$[3]
No Dif Modelled A Conditions

	θ_{dis}=0.0						θ_{dis}=-1.0					
	c=.00			c=.20			c=.00			c=.20		
	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50
a=0.5	.0203[1] .0219[2] -8%[3]	.0206 .0206 .05%	.0210 .0195 8%	.0200 .0203 -2%	.0194 .0197 -2%	.0191 .0198 -4%	.0245 .0212 16%	.0258 .0251 3%	.0270 .0308 3%	.0226 .0223 3%	.0224 .0225 -1%	.0223 .0247 -9%
a=1.0	.0262 .0282 -7%	.0267 .0240 11%	.0280 .0231 21%	.0236 .0220 8%	.0222 .0234 6%	.0217 .0208 4%	.0324 .0389 17%	.0364 .0505 -28%	.0406 .0368 10%	.0252 .0216 8%	.0250 .0243 3%	.0251 .0272 8%
a=1.5	.0318 .0350 -9%	.0330 .0306 8%	.0352 .0348 1%	.0265 .0234 13%	.0245 .0246 -.4%	.0237 .0221 7%	.0400 .0450 -11%	.0474 .0390 21%	.0553 .0656 -16%	.0272 .0292 -7%	.0268 .0281 -5%	.0270 .0286 -6%

Parameter values above are for both Focal and Reference Groups.

TABLE 7

$Var_{ht}ln(\alpha)$[1]
$Var_{emp}ln(\alpha)$[2]
$(Var_{ht}ln(\alpha)/Var_{emp}ln(\alpha))-1$[3]
DIF Modelled B Conditions

	θ_{dis}=0.0						θ_{dis}=-1.0					
	c=.00			c=.20			c=.00			c=.20		
	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50	b=.00	b=.30	b=.50
a=0.5	.0220[1] .0214[2] 3%[3]	.0226 .0218 4%	.0229 .0243 -5%	.0214 .0206 4%	.0210 .0202 4%	.0209 .0224 -7%	.0247 .0220 12%	.0253 .0250 1%	.0260 .0268 -3%	.0239 .0276 -13%	.0235 .0241 -2%	.0233 .0229 2%
a=1.0	.0250 .0239 5%	.0264 .0224 18%	.0281 .0272 3%	N/A	.0210 .0227 -7%	.0227 .0252 -10%	.0286 .0314 -9%	.0318 .0333 -5%	.0347 .0335 4%	N/A	.0248 .0263 -6%	.0246 .0178 38%
a=1.5	.0270 .0217 24%	.0295 .0311 -5%	.0322 .0367 -12%	.0247 .0255 -3%	.0241 .0227 6%	.0239 .0179 21%	.0320 .0300 7%	.0380 .0368 3%	.0447 .0453 -1%	.0260 .0274 -5%	.0255 .0258 -1%	.0253 .0252 3%

Parameter values above are for Focal Groups.
Reference Group ICCs are (1, 0,-.2).

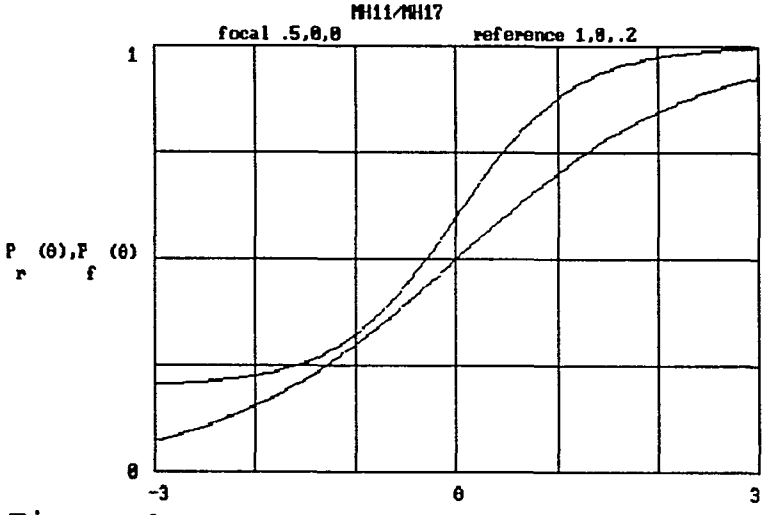

Figure 1

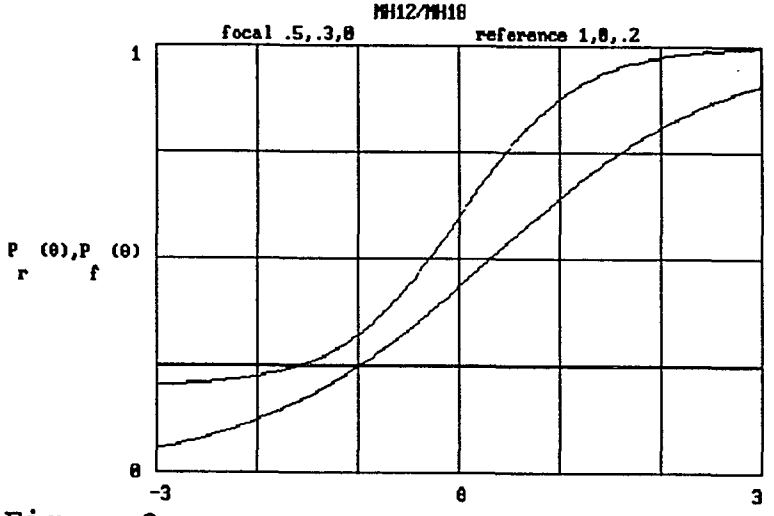

Figure 2

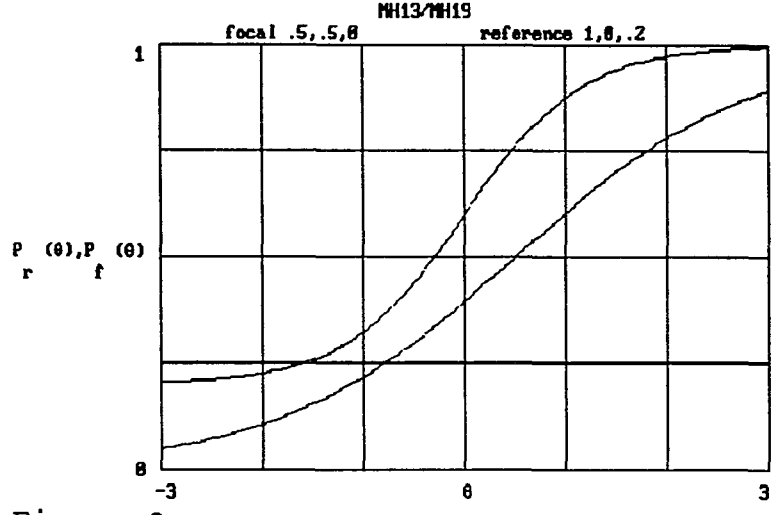

Figure 3

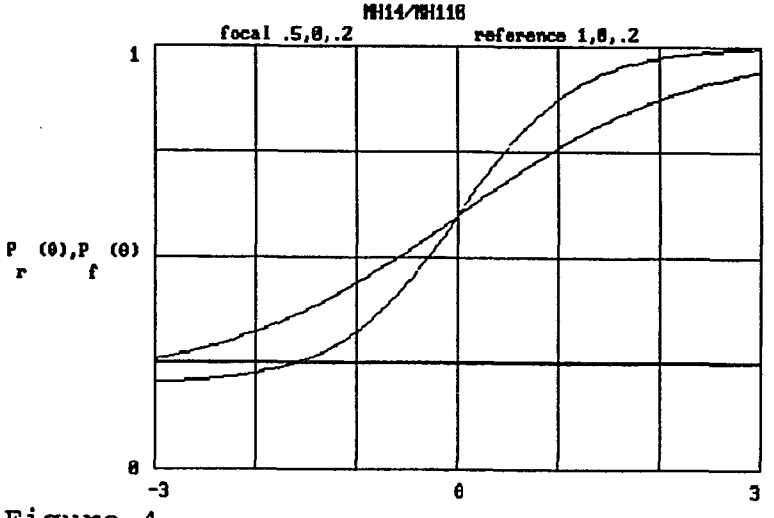

Figure 4

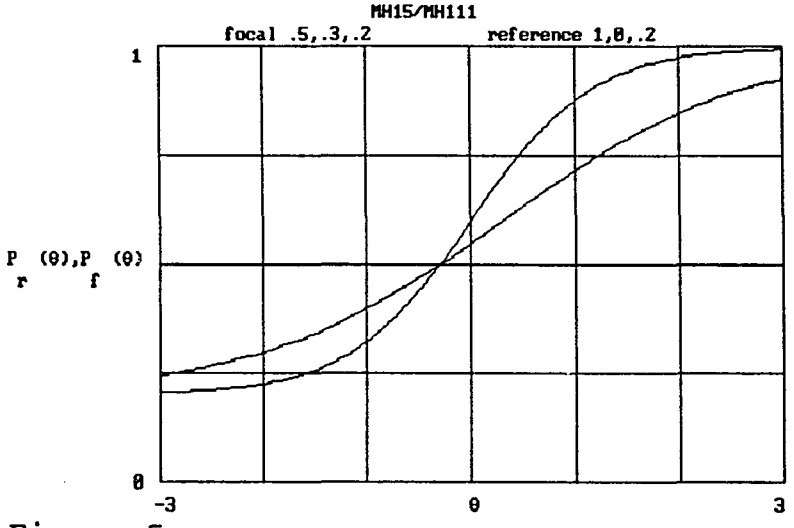

Figure 5

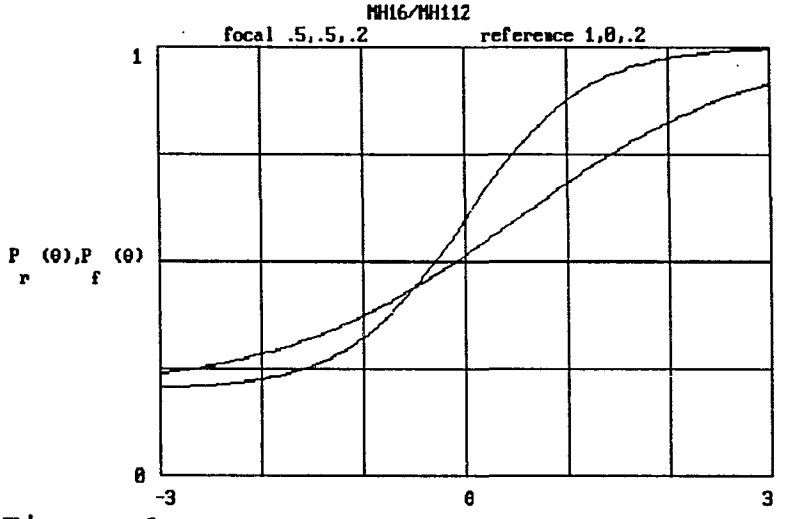

Figure 6

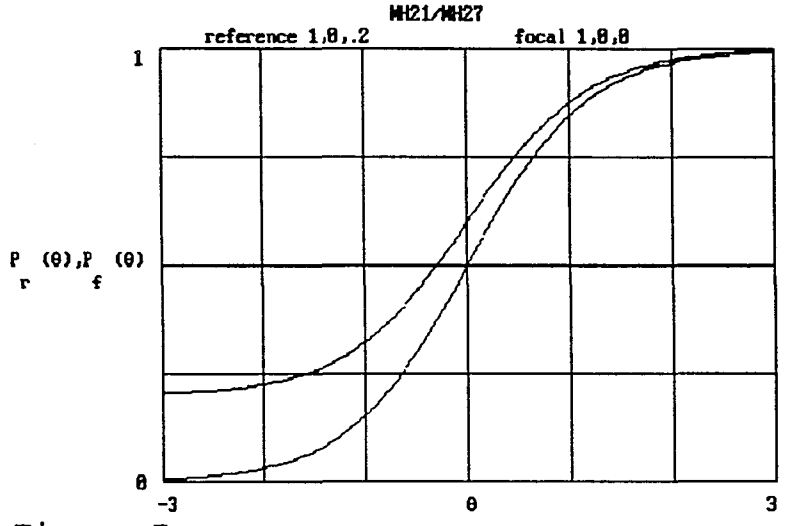

Figure 7

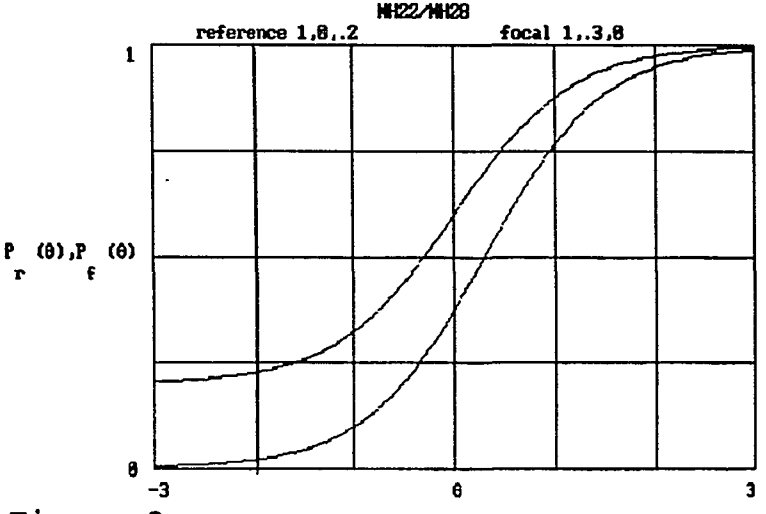

Figure 8

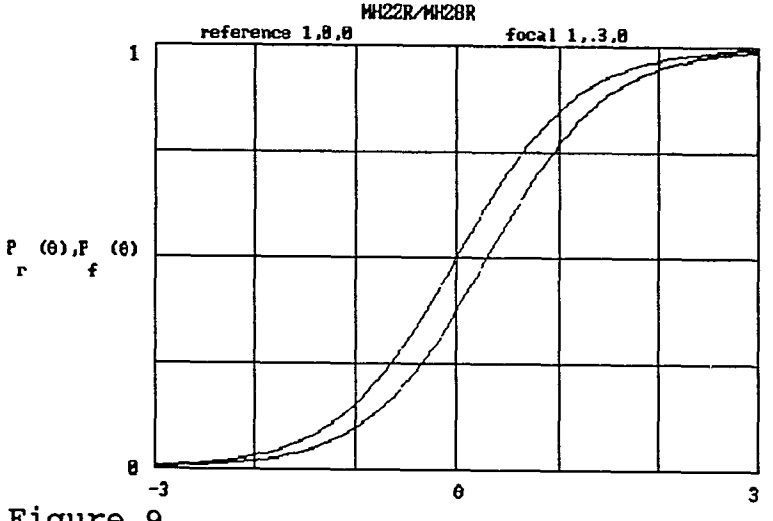

Figure 9

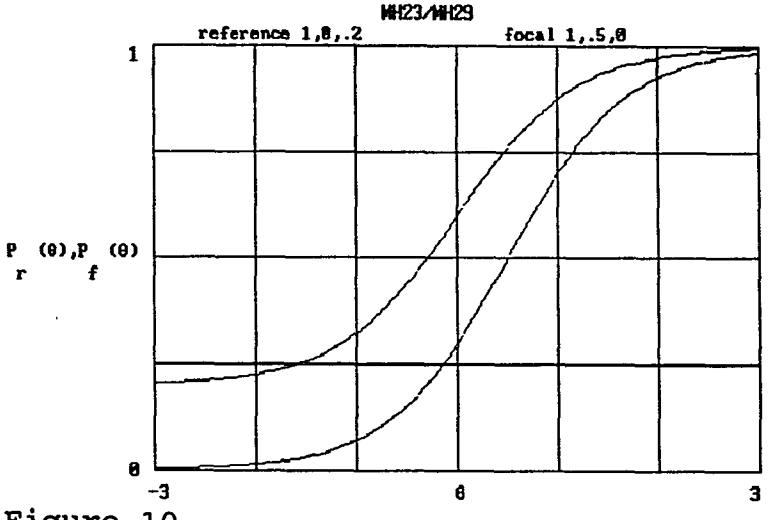

Figure 10

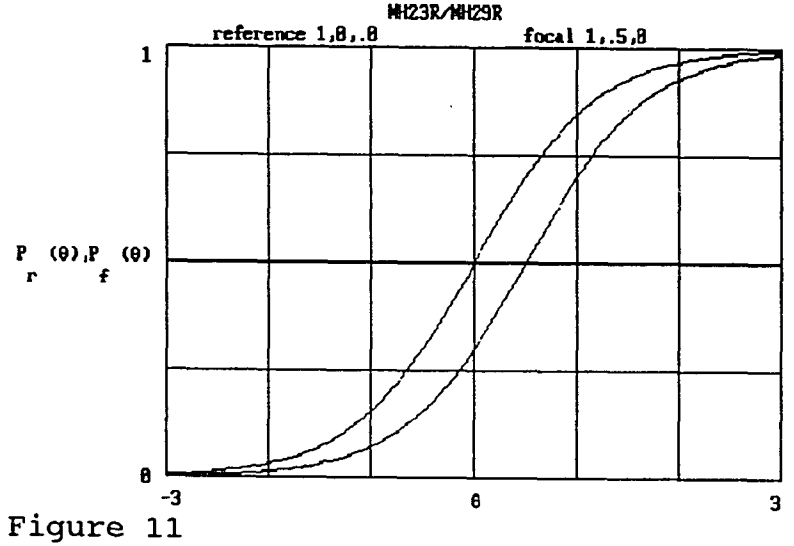

Figure 11

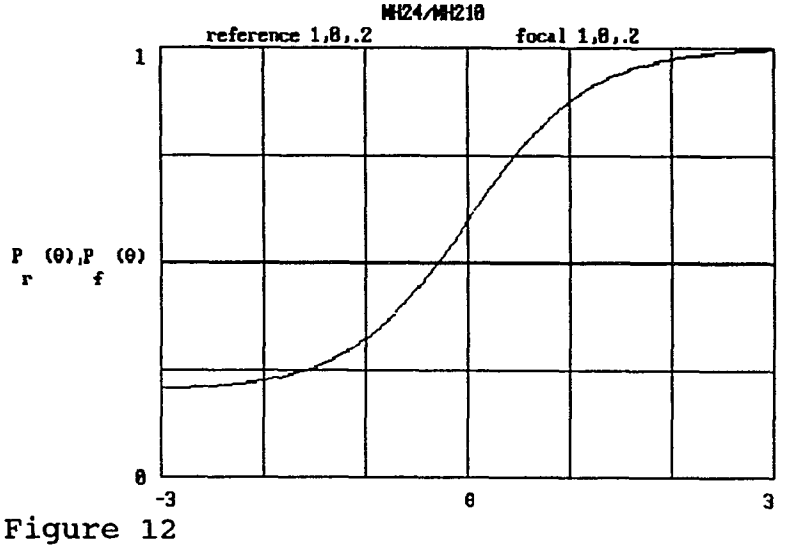

Figure 12

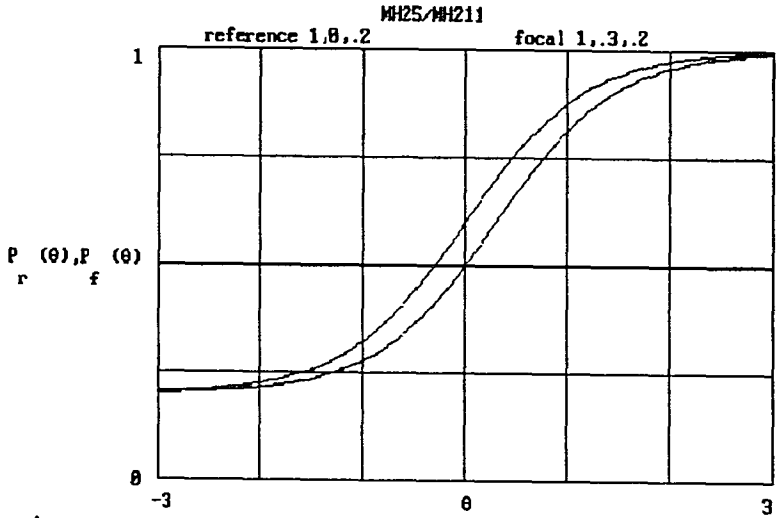

Figure 13

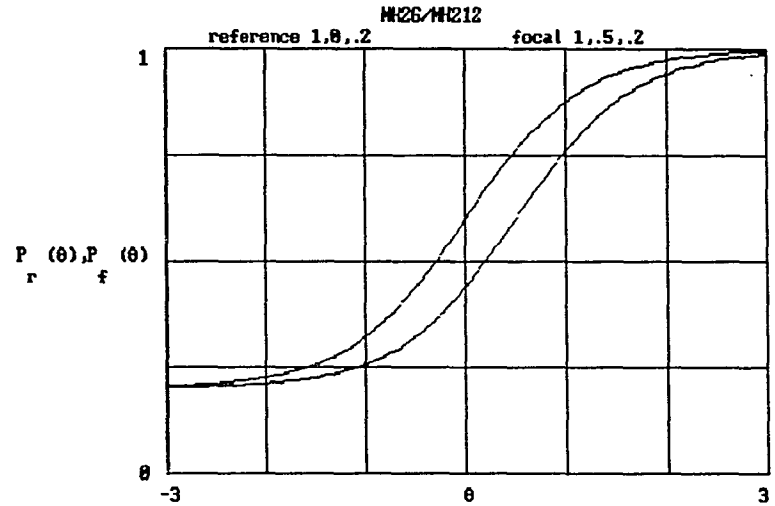

Figure 14

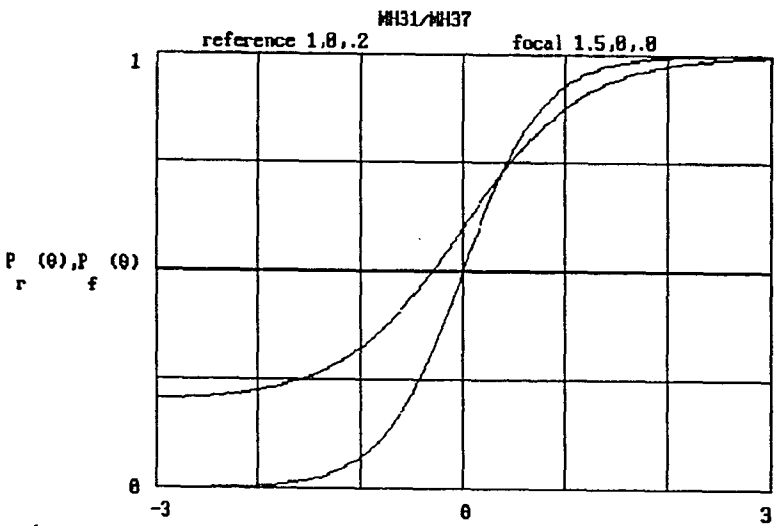

Figure 15

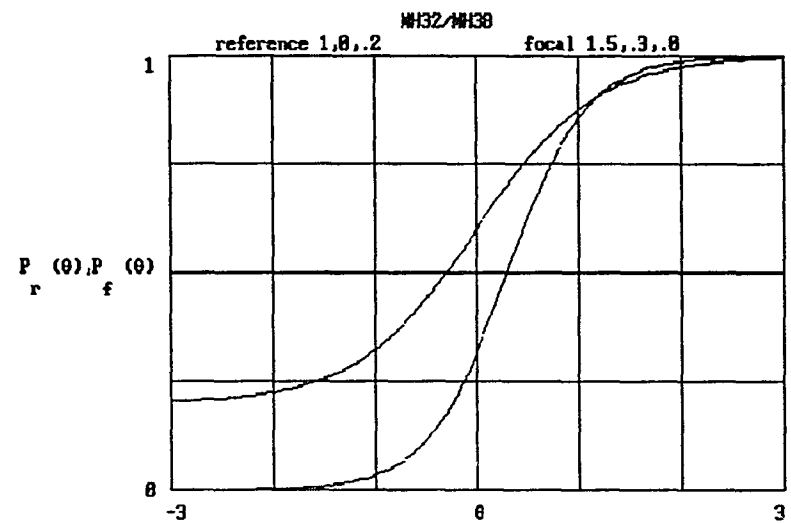

Figure 16

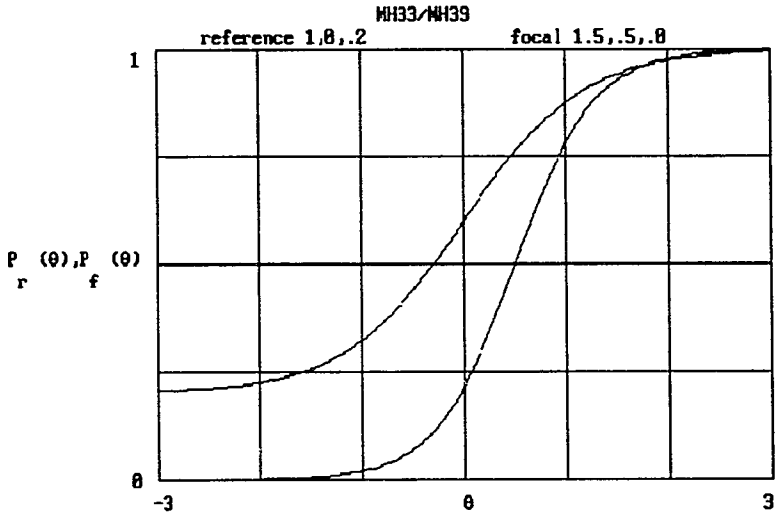

Figure 17

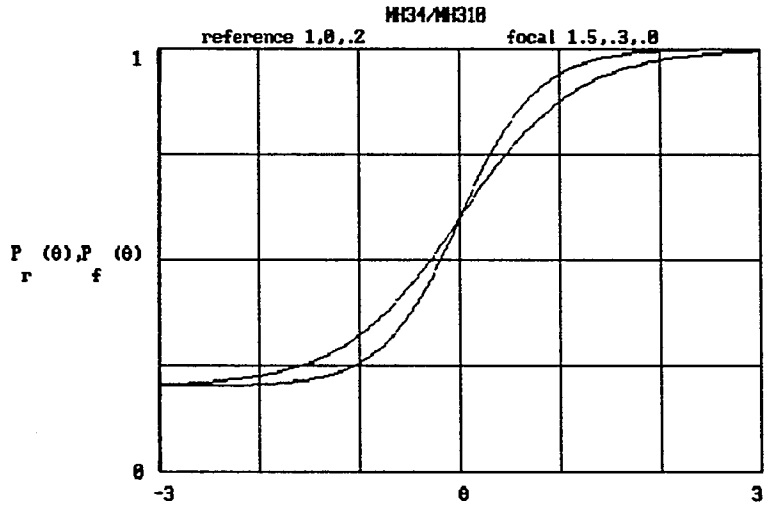

Figure 18

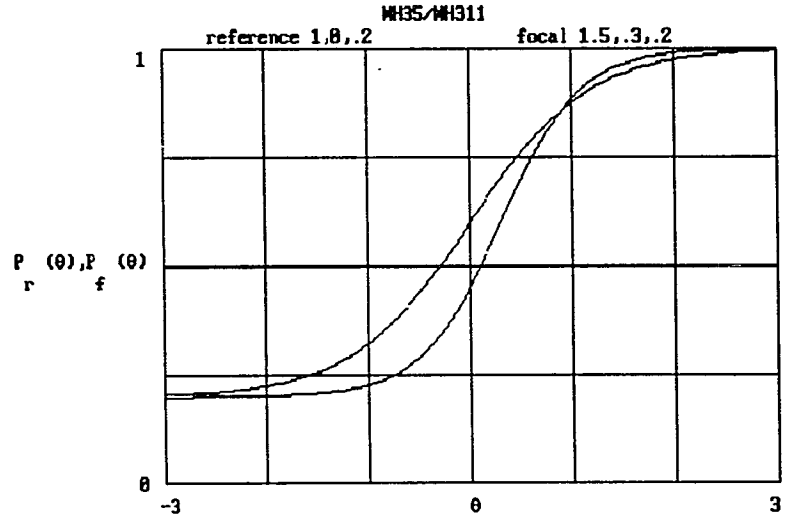

Figure 19

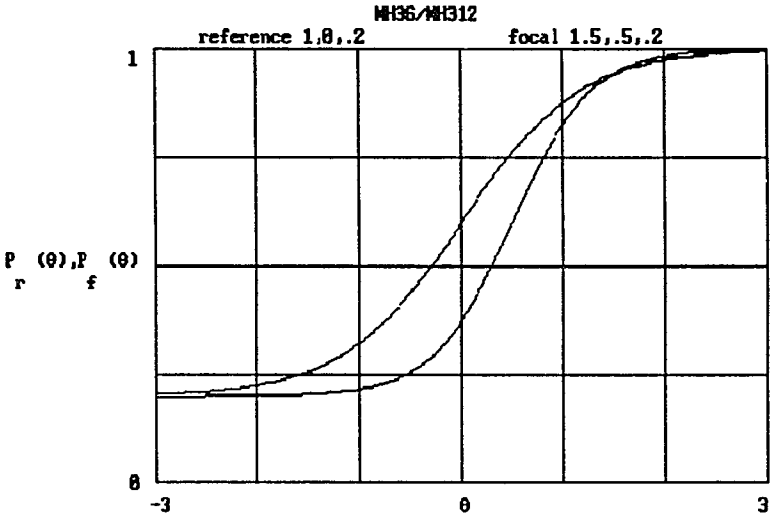

Figure 20

position (2,4) of the tables. These rows and columns correspond to combinations of ICC parameters for the focal groups in the type I cases and the focal and reference groups in the type II cases. Cases which differ only on the Θ_{dis} factor have identical ICCs, eg. MH11 and MH17.

Type I Cases

ANOVA for MH_{CHISQ} type I error rate cases (Table 8) illustrates trends in the individual simulation data and appears to confirm that the level of Θ_{dis} and the manner in which it interacts with certain item parameters powerfully affects the level of the type I error rate for MH_{CHISQ}. The results obtained show a powerful main effect for Θ_{dis} with its associated F=59.070, statistically significant at the level of .002.

This implies that the MH procedure is broadly mistaking the modelled ability discrepancy for DIF under no DIF conditions. The overall mean type I error rate is .04 for Θ_{dis}=0 and .09 for Θ_{dis}=1. Since this is a main effect it is interpreted that Θ_{dis} affects the type I error rate over all levels of the item parameter variables. An examination of Table 2 shows that the presence of Θ_{dis} is associated with an increase in type I error rate in 16 out of the 18 sets of cases with identical item parameter values across the Θ_{dis} factor.

Statistical significance was also obtained for the A by C item parameter interaction with F=16.253, p =.012. The

TABLE 8

ANOVA Type I Error Rate

by A, B, C and Θ_{dis}

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.025	6	.004	11.587	.016
A	.002	2	.001	2.794	.174
B	.000	2	.000	.289	.794
C	.002	1	.002	4.287	.107
Θ_{dis}	.021	1	.021	59.070	.002
2 Way Interactions	.019	13	.001	4.095	.092
A B	.001	4	.000	.845	.563
A C	.012	2	.006	16.253	.012
A Θ_{dis}	.003	2	.001	3.635	.125
B C	.002	2	.001	2.226	.224
B Θ_{dis}	.000	2	.000	.651	.569
C Θ_{dis}	.002	1	.002	4.287	.107
3 Way Interactions	.011	12	.001	2.510	.194
A B C	.001	4	.000	.582	.694
A B Θ_{dis}	.000	4	.000	.258	.891
A C Θ_{dis}	.009	2	.005	12.745	.018
B C Θ_{dis}	.000	2	.000	.637	.575
Explained	.055	31	.002	4.932	.065
Residual	.001	4	.000		
Total	.057	35	.002		

summary means across these factors are given in Table 9.

TABLE 9

A by C Interaction-MH_{CHISQ} Type I

	c=.00	c=.20
a=0.5	.048	.078
a=1.0	.050	.052
a=1.5	.103	.043

These means reveal a trend of a disordinal interaction between the levels of A and C on the level of the type I error rate. Higher type I error rates are evident for low A parameter values with higher C parameter values and higher A parameter values with low C parameter values. Lower type I error rates are obtained for cases with low A parameter values with low C parameter values and high A parameter values with high C parameter values.

The A by C interaction is not constant across the levels of Θ_{dis}, however, and must be qualified further. Significance is obtained for the A by C by Θ_{dis} interaction with F=12.745, p =.018. Therefore the A by C interaction is not generalizable across the levels of the Θ_{dis} factor. The A by C

by Θ_{dis} interaction is of considerable practical and statistical significance. The effect is most conspicuous for cells in the half of the table with Θ_{dis}=-1.0 and the summary means are given in Table 10.

TABLE 10

A by C by Θ_{dis} Interaction-MH_{CHISQ} Type I

	Θ_{dis}=0		Θ_{dis}=-1	
	c=.00	c=.20	c=.00	c=.20
a=0.5	.040	.042	.057	.115
a=1.0	.037	.043	.063	.060
a=1.5	.042	.033	.163	.053

Type I error rate levels are higher than the .05 level in the half of the table with Θ_{dis}=-1.0 whereas for the half with Θ_{dis}=0.0 all type I error rates are below .05. The cells which deviate most sharply from the .05 significance level in the above table are located in the (1,4) and (3,3) positions.

Similar findings obtain for the ANOVA analyses for the α in the type I analyses which are given in Table 11. Significance was obtained for main effects A (F=129.536, p=.000), C (F=228.571, p=.000) and Θ_{dis} (F=24.143, p=.008) and two way interactions A by C (F=31.536, p=.004), A by Θ_{dis} (F=127.964, p=.000) and C by Θ_{dis} (F=180.036, p=.000). The

TABLE 11

ANOVA α by A, B, C and Θ_{dis}

Type II Cases

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.162	6	.027	86.690	.000
A	.081	2	.040	129.536	.000
B	.003	2	.001	4.179	.105
C	.071	1	.071	228.571	.000
Θ_{dis}	.008	1	.008	24.143	.008
2 Way Interactions	.159	13	.012	39.305	.001
A B	.002	4	.000	1.286	.407
A C	.020	2	.010	31.536	.004
A Θ_{dis}	.080	2	.040	127.964	.000
B C	.001	2	.000	1.000	.444
B Θ_{dis}	.001	2	.001	2.393	.207
C Θ_{dis}	.056	1	.056	180.036	.000
3 Way Interactions	.025	12	.002	6.714	.040
A B C	.001	4	.000	.839	.565
A B Θ_{dis}	.000	4	.000	.304	.863
A C Θ_{dis}	.022	2	.011	34.750	.003
B C Θ_{dis}	.002	2	.001	3.250	.145
Explained	.346	31	.011	35.861	.002
Residual	.001	4	.000		
Total	.347	35	.010		

most generalizable finding, as for MH_{CHISQ}, A by C by Θ_{dis} interaction (F=34.750, p=.003). The summary means are presented in Table 12.

TABLE 12

A by C by Θ_{dis} Interaction-α Type I

	Θ_{dis}=0		Θ_{dis}=-1	
	c=.00	c=.20	c=.00	c=.20
a=0.5	1.00	1.00	.95	.89
a=1.0	1.00	1.00	1.13	.98
a=1.5	1.01	1.01	1.30	1.00

All of the α values corresponding to Θ_{dis}=0 are quite close to 1.00 accuratly reflecting the no DIF condition. Howeever, for the half with Θ_{dis}=0.0 the α values deviate considerably from 1.00. all type I error rates are below .05. The cells which deviate most sharply are located in the (1,4) and (3,3) positions as in the MH_{CHISQ} Type I analyses.

The corresponding cases from Tables 2, 3, 4, 5, 6 and 7 are MH110A, MH111A and MH112A for the (1,4) position and MH37A, MH38A and MH39A for the (3,3) position. The cases are considered together as per their positions in the tables.

Cases MH110A (.5, 0, .2, -1), MH111A (.5, .3, .2, -1) and MH112A (.5, .5, .2, -1) are based on focal and reference group ICC parameter values of which differ only on the b parameter

values. The underlying ICCs are graphed in Figures 4, 5 and 6. It is noted that the ICCs for both groups are increasing broadly and gradually due to the low a and high c parameter values for each of the cases. These types of ICCs, which were used to generate the data, do not typically provide high levels of information as there is only moderate differentiation of p values over the Θ scale. In addition, the focal group is distributed below the reference groups on the Θ dimension in these cases. The MH procedure appears to be mistaking Θ_{dis} for DIF favoring the focal group, resulting in higher MH_{CHISQ} type I error rates.

The ETS DIF classifications for these cases show correct "A" classifications in proportion to the MH chi square error and that when items are misclassified it is consistently as a "B" item classification. $Var_{ht}ln(\alpha)$ and $Var_{emp}ln(\alpha)$ were both moderate in terms of the general trends and levels of the variances. $Var_{ht}ln(\alpha)$ approximated $Var_{emp}ln(\alpha)$ within 3%, -1%, and -9% respectively for these cases.

Cases MH37A (1.5, 0, 0, -1), MH38A (1.5, .3, 0, -1) and MH39A (1.5, .5, 0, -1) are based on focal and reference group ICC parameter values which again differ only on the b parameter values. The underlying ICCs are graphed in Figures 15, 16 and 17. It is noted that the ICCs for both groups in these figures are increasing narrowly and sharply for Θ between -1.0 and 1.0 due to high a and low c parameter values. Thus, there is a low probability of lower ability examinees

passing the item, especially focal group examinees who are shifted further to the right on the θ scale. This results in a low information in the data sets generated from these ICCs and causes the high MH_{CHISQ} type I error rates.

The ETS DIF classifications for these cases show correct "A" classifications in proportion to the MH chi square error rate. However, there is considerable inconsistency in the misclassification of items. The items are frequently misclassified as "C" items for these simulations. This is most likely due to the high levels of the variances for these cases. They are, in fact, the highest of the type I simulations. The discrepancies between $Var_{ht}\ln(\alpha)$ and $Var_{emp}\ln(\alpha)$ are also substantial in terms of absolute value at -11%, 21% and -16%.

Type II Cases

ANOVA for MH_{CHISQ} type II error rate cases (see Table 13) illustrates trends in the individual simulation data within the context of the focal group ICC varied as per the design with the reference group ICC fixed at (1, 0, .20). This results in two missing cells in the design (MH24/210) as the focal and reference group ICCs are identical in these cases. Thus, there is a degree of non-orthoganality which prevents estimation of 3-way interactions. Significance of the main effects is obtained for the A (F=4.736, p=.027) B (F=7.038, p=.008) and C (F=56.561, p=.000) item parameter factors. For the three A parameter factor values (0.5, 1.0, 1.5) the mean

TABLE 13

ANOVA Type II Error Rate

by A, B, C and Θ_{dis}

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	2.918	6	.486	13.534	.000
A	.340	2	.170	4.736	.027
B	.506	2	.253	7.053	.008
C	2.032	1	2.032	56.561	.000
Θ_{dis}	.050	1	.050	1.394	.257
2 Way Interactions	.630	13	.048	1.348	.293
A B	.092	4	.023	.641	.642
A C	.136	2	.068	1.894	.187
A Θ_{dis}	.028	2	.014	.386	.687
B C	.239	2	.119	3.325	.066
B Θ_{dis}	.034	2	.017	.475	.632
C Θ_{dis}	.013	1	.013	.357	.560
Explained	3.547	19	.187	5.196	.001
Residual	.503	14	.036		
Total	4.051	33	.123		

MH_{CHISQ} type II error rates are .42, .14 and .21 respectively. For the three B parameter factor values (0, .3, .5) they are .41, .27 and .14. For the two C parameter factor values (0, .20) they are .04 and .52. The θ_{dis} factor and the interaction effects are statistically insignificant. However, the B by C interaction approaches statistical significance (F=3.325, p=.066).

The ANOVA results for α in the type II cases are given in Table 14. Again, significant item parameter factor main effects are noted for A (F=19.50, p=.000), B (F=18.347, p=.000) and C (F=85.898, p=.000). For the levels of A the mean values of α are 1.59, 2.49 and 2.75, for the levels of B they are (1.56, 2.23 and 3.00) and for the levels of C they are 3.15 and 1.19. Significance is also obtained for the A by C (F=9.22, p=.003) and B by C (F=7.235, p=.007) interactions.

The ICC parameter factors appear to be significant for MH_{CHISQ} and α because they impact on the presence and degree of nonuniform DIF in most of the cases. The significant trend across the levels of the A factor is noted by examining the columns of the design. In 8 of the 12 columns the type II error rate becomes smaller with increasing levels of the A factor. Increasing the level of the A factor primarily influences the presence and degree of nonuniform DIF in this study. This is apparent in the cases in columns 5 and 11. MH15B/MH111B, MH25B/MH211B and MH35B/MH311B are graphed in Figures 5, 13 and 19 and differ only on the A factor and show

TABLE 14

ANOVA α by A, B, C and Θ_{dis}

Type II Cases

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	44.564	6	7.427	26.388	.000
A	10.980	2	5.490	19.505	.000
B	10.329	2	5.164	18.347	.000
C	24.178	1	24.178	85.898	.000
Θ_{dis}	.524	1	.524	1.861	.194
2 Way Interactions	17.062	13	1.312	4.663	.004
A B	3.401	4	.850	3.021	.054
A C	5.192	2	2.596	9.223	.003
A Θ_{dis}	2.240	2	1.120	3.980	.043
B C	4.073	2	2.036	7.235	.007
B Θ_{dis}	.156	2	.078	.277	.762
C Θ_{dis}	1.649	1	1.649	5.859	.030
Explained	70.777	19	3.725	13.234	.000
Residual	3.941	14	.281		
Total	74.718	33	2.264		

diminishing type II error rates. There also is a clear trend in the direction of increasing magnitude of both $Var_{ht}ln(\alpha)$ and $Var_{emp}ln(\alpha)$ across the levels of the A factor. A complete report of the variance results is given later.

For the B parameter factor there is a clear trend in the direction of lower levels of type II error rates for increasing levels of b across the other factors. Within the levels of the A and C factors increasing the level of the B factor tends to increase uniform DIF and reduce non-uniform DIF where potentially present.

For uniform DIF this trend is evident in cases MH11B/MH17B, MH12B/MH18B and MH13B/MH31B in Figures 1, 2 and 3 and cases MH25B/MH211B and MH26B/MH121B in Figures 13 and 14. For non-uniform DIF cases MH34B/MH310B, MH35B/MH311B, MH36B/MH312B in Figures 18, 19 and 20 evidence this trend.

The C factor ICC parameter main effect is the largest in magnitude of the three factors. At any given level of the A and B factors there are very large differences in the type I error rates across the levels of C, with consistently higher levels for c=.20. This is true for both the uniform and non-uniform cases. For uniform DIF, cases MH22B/MH28B are contrasted with MH25B/MH211B in Figures 8 and 13. The difference of c parameters in MH22B/MH28B provide enough information in the lower end of the θ scale for the type I error rate to be quite low. However, in MH25B/MH211B the equivalence of c values limits the amount of DIF to a small

band over most of the θ scale. It is noted that performance is especially poor in case MH211B and the focal group θ distribution is shifted largely to the range of little DIF.

The effect of the C factor in nonuniform DIF cases is evident in MH32B/MH38B and MH35B/MH311B which are graphed in Figures 16 and 19. For MH32B/MH38B the difference between the focal and reference group c parameters again causes substantial DIF in the lower range of the θ which is easily detected by the procedure despite the small degree of nonuniform DIF in the upper end of the θ scale.

For MH35B/MH311B identical c parameters for the focal and reference groups again reduces the DIF in the lower range of the θ scale. This, along with the nonuniform DIF in the upper range of the θ scale causes the increased type I error rates in these cases.

The C factor also has significant effects on both $Var_{ht}ln(\theta)$ and $Var_{emp}ln(\theta)$ which are fully reported later.

An overall trend from this facet of the study is that the item parameter factors substantially effect the type II error rate through their effects on the uniformity of DIF. A t-test was run to assess if mean type II error rates differed across two groups of cases defined in terms of uniform and nonuniform DIF. The results were significant ($T=-2.42$, $p=.021$).

Phase II-Rasch Cases

The results of this section are presented in Table 15 outlining the seven simulated cases studied. Estimates were

obtained for the MH statistics under study for the studied item. The cases were designed so as to provide increasing overall discrepancies between the location of the b parameters of the items in the matching criteria M, the location of the mean θ for the focal and reference groups and the location of the studied items as per their b parameter values. The location of the studied items remain constant with the reference group ICC (1, 0, 0) and the focal group ICC (1, .3, 0). The difficulties of the items (b parameters) and the abilities of the subjects (θ) are on the same scale.

It is noted that the level of α remains relatively constant and stable over the seven cases at approximately 1.70 regardless of the increasingly discrepant relationship between the difficulties of the studied item for the focal and reference groups and both 1) the θ levels of the focal and reference groups and 2) the increasingly extreme range of the difficulties of the studied items in M. Moving the θ levels of subjects in an increasingly negative direction and the range of item difficulties in M in an increasingly positive direction results in lower probabilities of most of the simulated examinees passing the item. This in turn results in a lower level of item information. This factor does not appear to substantially effect the level of α. The level of the ETS Δ_{mh} statistic is also not strongly affected by the increasingly discrepant simulation conditions. All of the estimates are at approximately -1.20. Overall, the α and

TABLE 15

Rasch Cases

Case	Mean θ_r	Mean θ_f	b parameter values of items in M by .25	Alpha	Delta	MH Chi Square	Type II Error Rate	$Var_{ht}\ln(\alpha)$	Empirical Variance of $\ln(\alpha)$	Var Discrepancy
1	0.0	0.0	-2.25 to 2.25	1.70	-1.21	10.24	.110	.026	.025	-4%
2	0.0	-1.0	-2.25 to 2.25	1.73	-1.26	8.56	.220	.035	.029	-20%
3	0.0	-1.0	-1.25 to 3.25	1.70	-1.21	8.09	.230	.036	.033	-9%
4	0.0	-1.0	-1.00 to 3.50	1.71	-1.22	8.17	.245	.036	.034	-6%
5	0.0	-1.0	-0.75 to 3.75	1.72	-1.23	8.22	.240	.036	.034	-6%
6	-0.5	-1.5	-0.75 to 3.75	1.72	-1.22	6.35	.340	.048	.053	9%
7	-0.5	-1.5	0.0 to 3.75[1]	1.69	-1.18	5.50	.410	.051	.046	-11%

1. Two items each were modelled for b=3.25, 3.50 and 3.75.

Studied Item ICCs-Reference Group (1, 0, 0) Focal Group (1, .3, 0).

Δ_{mh} estimates appear to be quite robust to discrepant simulated conditions. This fortunate situation does not obtain with the other statistics under study. The mean MH Chi Square does not remain stable and becomes increasingly smaller over the seven simulated cases. There are only slightly more than half as many significant chi squares for case 7 compared to case 1, which might be considered an ideal case. This situation is reflected in the type II error rate which dramatically increases along with the increasingly more discrepant conditions. For case 1 we obtain a type II error rate of .11 whereas for case 7 that rate has reached .41.

The variance measure $Var_{ht}ln(\alpha)$ and the empirical variance of $ln(\alpha)$ both increase notably over the seven simulated conditions. The case 7 variances are larger by a factor of approximately 2 than the case 1 variances. There are differences between the estimated variances, $Var_{ht}ln(\alpha)$, and the empirical variances in each of the cases studied. In terms of absolute magnitude, these range from a -4% discrepancy for case 1 to a -20% discrepancy for case 2. No clear cut pattern of these magnitude inconsistencies is readily detected. However, it is noted that in six of the seven cases the estimated variance, $Var_{ht}ln(\alpha)$, overestimates the variance calculated empirically.

Phase III-Variance Analyses

Six ANOVA analyses were performed in an attempt to discern influences affecting the empirical variance, $Var_{ht}ln(\alpha)$ and the

discrepancy between these variances among the item parameter and Θ_{dis} factors. Since there is one case per cell no effort was made to evaluate the highest or fourth order interaction.

The ANOVA results for the empirical variance in the type I error rate condition are given in Table 16. Significance was obtained for the A parameter (F=16.841, p=.011), C parameter (F=32.741, sig.=.005) and Θ_{dis} (F=26.556, p=.007) main effects. Significance was also obtained for the C by Θ_{dis} interaction (F=8.047, p=.047).

The ANOVA results for $Var_{ht}\ln(\alpha)$ in the type I error rate case are presented in table 17. Most of the main effect and factor interactions are significant in this analysis. Significance was obtained for all of the main effects; A parameter (F=358.593, p=.000), B parameter (F=13.541, p=.017), C parameter (F=587.627, p=.000) and Θ_{dis} (F=322.102, p=.000). Among the two way interactions the following were significant: A by C (F=109.335, p=.000), A by Θ_{dis} (F=12.394, p=.019), B by C (F=29.252, p=.004), B by Θ_{dis} (F=12.742, p=.018) and C by Θ_{dis} (F=109.004, p=.000). Among the three way interactions significance was obtained for A by C by Θ_{dis} (F=17.909, sig.=.010).

The ANOVA results for the type II error rate cases are presented in Tables 18 and 19. It is important to note that the focal and reference group item parameters differ in the type II error rate cases. The parameter factors analyzed in the ANOVA are for the focal group but the results are also

TABLE 16

ANOVA Empirical Variance

by A, B, C and Θ_{dis}

Type I Case

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.002	6	.000	15.703	.009
A	.001	2	.000	16.841	.011
B	.000	2	.000	.619	.583
C	.001	1	.001	32.741	.005
Θ_{dis}	.001	1	.001	26.556	.007
2 Way Interactions	.001	13	.000	2.629	.181
A B	.000	4	.000	1.257	.415
A C	.000	2	.000	6.668	.053
A Θ_{dis}	.000	2	.000	1.571	.314
B C	.000	2	.000	.425	.680
B Θ_{dis}	.000	2	.000	1.866	.268
C Θ_{dis}	.000	1	.000	8.047	.047
3 Way Interactions	.000	12	.000	1.181	.478
A B C	.000	4	.000	1.692	.311
A B Θ_{dis}	.000	4	.000	.746	.608
A C Θ_{dis}	.000	2	.000	1.516	.324
B C Θ_{dis}	.000	2	.000	.694	.551
Explained	.003	31	.000	4.599	.073
Residual	.000	4	.000		
Total	.003	35	.000		

TABLE 17

ANOVA $Var_{ht}\ln(\alpha)$

by A, B, C and Θ_{dis}

Type I Case

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.002	6	.000	275.666	.000
A	.001	2	.000	358.593	.000
B	.000	2	.000	13.541	.017
C	.000	1	.000	587.627	.000
Θ_{dis}	.000	1	.000	322.102	.000
2 Way Interactions	.000	13	.000	34.464	.002
A B	.000	4	.000	2.894	.164
A C	.000	2	.000	109.335	.000
A Θ_{dis}	.000	2	.000	12.394	.019
B C	.000	2	.000	29.252	.004
B Θ_{dis}	.000	2	.000	12.742	.018
C Θ_{dis}	.000	1	.000	109.004	.000
3 Way Interactions	.000	12	.000	6.005	.049
A B C	.000	4	.000	4.545	.086
A B Θ_{dis}	.000	4	.000	2.217	.230
A C Θ_{dis}	.000	2	.000	17.909	.010
B C Θ_{dis}	.000	2	.000	4.594	.092
Explained	.002	31	.000	70.132	.000
Residual	.000	4	.000		
Total	.002	35	.000		

TABLE 18

ANOVA Empirical Variance

by A, B, C and Θ_{dis}

Type II Case

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.002	6	.000	15.703	.009
A	.001	2	.000	16.841	.011
B	.000	2	.000	.619	.583
C	.001	1	.001	32.741	.005
Θ_{dis}	.001	1	.001	26.556	.007
2 Way Interactions	.001	13	.000	2.629	.181
A B	.000	4	.000	1.257	.415
A C	.000	2	.000	6.668	.053
A Θ_{dis}	.000	2	.000	1.571	.314
B C	.000	2	.000	.425	.680
B Θ_{dis}	.000	2	.000	1.866	.268
C Θ_{dis}	.000	1	.000	8.047	.047
3 Way Interactions	.000	12	.000	1.181	.478
A B C	.000	4	.000	1.692	.311
A B Θ_{dis}	.000	4	.000	.746	.608
A C Θ_{dis}	.000	2	.000	1.516	.324
B C Θ_{dis}	.000	2	.000	.694	.551
Explained	.003	31	.000	4.599	.073
Residual	.000	4	.000		
Total	.003	35	.000		

TABLE 19

ANOVA $Var_{ht}ln(\alpha)$

by A, B, C and Θ_{dis}

Type II Case

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.001	6	.000	129.918	.000
A	.000	2	.000	141.572	.000
B	.000	2	.000	16.147	.012
C	.000	1	.000	306.512	.000
Θ_{dis}	.000	1	.000	157.558	.000
2 Way Interactions	.000	13	.000	17.021	.007
A B	.000	4	.000	4.513	.087
A C	.000	2	.000	49.105	.002
A Θ_{dis}	.000	2	.000	5.020	.081
B C	.000	2	.000	27.976	.004
B Θ_{dis}	.000	2	.000	3.255	.145
C Θ_{dis}	.000	1	.000	32.515	.005
3 Way Interactions	.000	12	.000	4.421	.082
A B C	.000	4	.000	5.005	.074
A B Θ_{dis}	.000	4	.000	1.217	.427
A C Θ_{dis}	.000	2	.000	11.109	.023
B C Θ_{dis}	.000	2	.000	2.972	.162
Explained	.001	31	.000	33.995	.002
Residual	.000	4	.000		
Total	.001	35	.000		

based on a reference group parameterization of (1, 0, .20).

The ANOVA for the empirical variance in the type II error rate condition is given in Table 18. Among the main effects significance was obtained for A parameter (F=25.288, p=.005) C parameter (F=69.665, p=.001) and Θ_{dis} factor (F=39.805, p=.003).

For the two way interactions significance was obtained for the A by C (F=16.301, p=.012), B by C (F=18.182, p=.010) and C by Θ_{dis} (F=8.662, p=.042) interactions. Significance was also obtained for the A by B by C interaction (F=6.391, p=.050).

The ANOVA for $Var_{ht}\ln(\alpha)$ in the type II error rate case is given in Table 19. Significance was obtained for all the main effects; A parameter (F=141.572, p=.000), B parameter (F=16.147, p=.012) C parameter (F=306.512, p=.000) and Θ_{dis} factor (F=157.558, p=.000). Among the two way interactions significance was obtained for the A by C (F=49.105, p=.002), B by C (F=27.976, p=.004) and C by Θ_{dis} (F=32.515, p=.005) interactions. Significance was also obtained for the A by C by Θ_{dis} interaction (F=11.109, p=.023).

ANOVA analyses for the discrepancies between $Var_{ht}\ln(\alpha)$ and the empirical variance in the type I and type II error rate cases were performed and are presented in Tables 20 and 21, respectively. Interestingly, statistical significance was not obtained for any of the main effect or interaction in either analyses.

TABLE 20

ANOVA Variance Discrepancy

by A, B, C and θ_{dis}

Type I Case

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.025	6	.004	.494	.790
A	.009	2	.005	.540	.620
B	.001	2	.001	.061	.942
C	.001	1	.001	.133	.734
θ_{dis}	.014	1	.014	1.630	.271
2 Way Interactions	.141	13	.011	1.302	.435
A B	.050	4	.012	1.487	.355
A C	.007	2	.003	.389	.701
A θ_{dis}	.022	2	.011	1.287	.370
B C	.022	2	.011	1.293	.369
B θ_{dis}	.035	2	.018	2.105	.237
C θ_{dis}	.007	1	.007	.831	.413
3 Way Interactions	.176	12	.015	1.752	.311
A B C	.081	4	.020	2.430	.205
A B θ_{dis}	.055	4	.014	1.646	.320
A C θ_{dis}	.010	2	.005	.593	.595
B C θ_{dis}	.030	2	.015	1.768	.282
Explained	.342	31	.011	1.320	.439
Residual	.033	4	.008		
Total	.375	35	.011		

TABLE 21

ANOVA Variance Discrepancy

by A, B, C and Θ_{dis}

Type II Case

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig. of F
Main Effects	.015	6	.003	.185	.966
A	.010	2	.005	.356	.721
B	.004	2	.002	.149	.866
C	.000	1	.000	.010	.926
Θ_{dis}	.001	1	.001	.089	.781
2 Way Interactions	.128	13	.010	.714	.712
A B	.023	4	.006	.425	.786
A C	.007	2	.003	.238	.799
A Θ_{dis}	.007	2	.004	.254	.787
B C	.047	2	.024	1.715	.290
B Θ_{dis}	.041	2	.021	1.489	.329
C Θ_{dis}	.003	1	.003	.193	.683
3 Way Interactions	.185	12	.015	1.114	.505
A B C	.067	4	.017	1.211	.429
A B Θ_{dis}	.046	4	.011	.831	.569
A C Θ_{dis}	.070	2	.035	2.540	.194
B C Θ_{dis}	.002	2	.001	.063	.940
Explained	.328	31	.011	.767	.710
Residual	.055	4	.014		
Total	.383	35	.011		

Phase IV-Eta Analyses

The cases analyzed in this phase are referred to by their row and column positions in tables 22 and 23 with the prefix EMH and suffix A or B for type I and type II error cases, respectively. For example EMH17A refers to the case in the (1,7) positions in the type I analyses in Table 22.

The eta analyses were accomplished in a two step process. First the level of eta was manipulated by varying the item parameters in the matching criteria M. The results of these analyses are presented for the type I and type II cases in Tables 22 and 23, respectively. Each table also indicates the level of eta for the focal and reference groups and the ETS DIF classifications. These were obtained for each set of conditions based on the item parameterizations of M and Θ_{dis} between the focal and reference groups.

The second step of the analyses was to perform multiple regression with the level of the type I and type II error rates as the dependent variables. Θ_{dis} was a classification factor, eta a continuous variable and eta by Θ_{dis} the interaction term. The results are summarized in Tables 24 and 25. Independent variables were entered individually to determine the incremental contribution of each to multiple R and R^2.

For the type I error case examination of Table 24 shows that each independent variable entered in the regression makes a significant contribution to multiple R and R^2. Entered in

TABLE 22

Eta[2] for Reference[1] and Focal[2] Groups
Type I Error Rate[3]
ETS A[4], B[5] and C[6] DIF Classifications

	θ_{dis}=0.0					θ_{dis}=-1.0				
	a=.25	a=.40	a=.75	a=1.25	a=1.75	a=.25	a=.40	a=.75	a=1.25	a=1.75
c=.00	.716[1] .719[2] .015[3] A-197[4] B- 3[5] C- 0[6]	.829 .831 .025 A-195 B- 5 C- 0	.917 .917 .03 A-194 B- 6 C- 0	.950 .951 .055 A-189 B- 11 C- 0	.963 .963 .055 A-189 B- 11 C- 0	.712 .701 .905 A- 19 B-119 C- 62	.830 .814 .41 A-118 B- 79 C- 3	.916 .901 .04 A-192 B- 8 C- 0	.951 .935 .03 A-194 B- 6 C- 0	.964 .947 .06 A-188 B- 12 C- 0
c=.10	.685 .685 .04 A-192 B- 8 C- 0	.800 .801 .035 A-193 B- 7 C- 0	.894 .894 .065 A-187 B- 13 C- 0	.929 .930 .025 A-195 B- 5 C- 0	.944 .944 .025 A-195 B- 5 C- 0	.682 .660 .985 A- 3 B- 92 C-105	.801 .773 .63 A- 74 B-109 C- 17	.893 .859 .135 A-173 B- 26 C- 1	.930 .891 .065 A-187 B- 13 C- 0	.943 .903 .045 A-191 B- 8 C- 1
c=.20	.649 .651 .035 A-193 B- 7 C- 0	.768 .768 .04 A-192 B- 8 C- 0	.868 .869 .05 A-190 B- 10 C- 0	.908 .909 .04 A-192 B- 8 C- 0	.923 .921 .005 A-199 B- 1 C- 0	.650 .621 .995 A- 1 B- 75 C-124	.770 .731 .72 A- 57 B-115 C-28	.870 .821 .190 A-162 B- 36 C- 2	.908 .853 .075 A-185 B- 14 C- 1	.922 .865 .025 A-195 B- 5 C- 0

TABLE 23

Eta[2] for Reference[1] and Focal[2] Groups
Type II Error Rate[3]
ETS A[4], B[5] and C[6] DIF Classifications

	θ_{dis}=0.0					θ_{dis}=-1.0				
	a=.25	a=.40	a=.75	a=1.25	a=1.75	a=.25	a=.40	a=.75	a=1.25	a=1.75
c=.00	.717[1] .717[2] .215[3] A- 43[4] B-148[5] C- 9[6]	.830 .831 .180 A- 36 B-144 C- 20	.917 .917 .165 A- 33 B-125 C- 42	.951 .951 .150 A- 30 B-128 C- 42	.963 .963 .175 A- 35 B-107 C- 58	.719 .695 .000 A- 0 B- 4 C-196	.829 .812 .020 A- 4 B- 45 C-151	.917 .901 .185 A- 37 B- 95 C- 68	.950 .934 .330 A- 66 B- 96 C- 38	.963 .948 .44 A- 88 B- 88 C- 24
c=.10	.685 .682 .185 A- 37 B-152 C- 11	.801 .799 .190 A- 38 B-130 C- 32	.894 .893 .170 A- 34 B-133 C- 33	.929 .930 .150 A- 30 B-138 C- 32	.944 .944 .125 A- 25 B-126 C- 49	.687 .652 .000 A- 0 B- 1 C-199	.801 .769 .000 A- 0 B- 20 c-180	.894 .856 .095 A- 19 B-102 C- 79	.930 .890 .205 A- 41 B-100 C- 59	.943 .902 .195 A- 39 B-116 C- 45
c=.20	.652 .649 .175 A- 35 B-155 C- 10	.768 .768 .190 A- 38 B-140 C- 22	.868 .867 .175 A- 35 B-138 C- 27	.908 .908 .150 A- 30 B-122 C- 48	.922 .922 .125 A- 25 B-134 C- 0	.652 .613 .000 A- 0 B- 0 C-200	.770 .727 .01 A- 2 B- 11 C-187	.870 .820 .020 A- 4 B- 71 C-125	.907 .852 .105 A- 21 B-100 C- 79	.923 .864 .160 A- 32 B- 99 C- 69

TABLE 24

Eta Analysis

Multiple Regression

Type I Error Rate-DV

Θ_{dis}, Eta and Θ_{dis} by Eta

Summary Table

Step	Mult R	R^2
1	.6226	.3876
2	.7806	.6094
3	.9815	.9633

Step 1.

Variable	F	Sig. F
Eta	17.721	.0002

Step 2.

Variable	F	Sig. F
Eta	23.516	.0000
Θ_{dis}	15.329	.0006

Step 3.

Variable	F	Sig. F
Eta	.021	.8848
Θ_{dis}	295.691	.0000
Θ_{dis} by Eta	250.810	.0000

TABLE 25

Eta Analysis

Multiple Regression

Type II Error Rate-DV

θ_{dis}, Eta and θ_{dis} by Eta

Summary Table

Step	Mult R	R^2
1	.4705	.2214
2	.5199	.2703
3	.8091	.6547

Step 1.

Variable	F	Sig. F
Eta	7.962	.0087

Step 2.

Variable	F	Sig. F
Eta	7.520	.0107
θ_{dis}	1.811	.1896

Step 3.	F	Sig. F
Eta	1.056	.3136
θ_{dis}	31.000	.0000
θ_{dis} by Eta	28.941	.0000

step I, the eta is highly significant and results in multiple R= .6226 and R^2=.3876. The entry of the Θ_{dis} factor in the second step is also highly significant and contributes to multiple R and R^2 with magnitudes of .7806 and .6094 respectively. The entry of the Θ_{dis} by eta interaction term in the third step is also highly significant bringing multiple R to .9815 and R^2 to .9633. It is important to note that at the third step the eta factor is no longer significant and the variation explained is almost entirely through the Θ_{dis} factor and the eta by Θ_{dis} interaction.

A somewhat similar pattern of results obtains in the type II error rate multiple regression analysis summarized in Table 24. At the first step eta is significant with multiple R=.4705 and R^2=.2214. In the second step Θ_{dis} is entered but is not significant, multiple R=.5199 and R^2=.2703. Upon entry of the Θ_{dis} by eta interaction term in the third step a pattern of results similar to the type I analysis is obtained. Θ_{dis} and Θ_{dis} by Eta are significant and Eta is no longer significant. For the third step Multiple R=.8091 and R^2=.6547.

Analysis of Table 22 indicates the general trends in the type I cases. For the half of the table with Θ_{dis}=0.0 we find type I error rates generally acceptable at the .05 level. In examining the Θ_{dis}=-1.0 portion of the table we note the high levels of type I error associated with particular ranges of mean eta. The cases in column 6, EMH16A, EMH26A and EMH36A

show particularly high type I error rates. With this particular range of eta and θ_{dis}=-1.0 the procedure is strongly indicating DIF when none exists. In column 7, cases EMH17A, EMH27A and EMH37A also evidence high levels of type I error. The individual cases of this portion of the table show that the type I error rate is strongly related to the mean level of eta and thus the eta by θ_{dis} interaction is implicated in the high type I error rates.

For the type II error rate cases outlined in Table 23 an analogous situation obtains. For the half of the table with θ_{dis}=0.0 the type II error rates are below .20 for all cases except EMH11B. Thus power is generally in excess of .80 which is often considered an acceptable level. For the portion of the table with θ_{dis}=-1.0 there is a relation between the level of eta and the type II error rate. the relation is the inverse of that in the type I cases, however. In the presence of the θ_{dis} factor higher levels of eta are associated with higher levels of the type II error rate. For example, cases EMH19B and EMH110B have the highest mean eta rates in Table 23 and also have the highest type II error rates.

V. Discussion

The results of this study demonstrate that there are some serious obstacles to the simple, direct application of the Mantel Haenszel procedure for DIF analyses. The specific issues raised are with regard to the accuracy of the procedure given certain item ICC parameterizations of studied items and items in the matching criteria and given discrepancies between focal and reference group θ levels. These concerns are especially relevant since the Mantel Haenszel procedure has been proposed as a flexible, inexpensive and accurate DIF detection technique.

This study proceeded from the premise that the IRT DIF definition (Lord, 1980) is theoretically preferred and that the MH procedure is intended to approximate IRT DIF. Thus, the data are generated from an IRT DIF model. For a DIF procedure to be useful in a general sense it should detect DIF under one, two and three parameter IRT models and with focal and reference groups that differ on the θ dimension.

This study evaluated the MH procedure in terms of statistical concepts such as type I and type II error rates, α_{mh} and Δ_{mh} estimators, variances and ETS DIF classifications. The MH procedure has generally been received with enthusiasm based on a limited number of comparison studies in which IRT methods and the MH procedure have found agreement (Hambleton & Rogers, 1989; Thissen Steinberg & Wainer, 1988; Harvey, 1990; Spray, 1990). However, the implications of theoretical

work (Zwick, 1990; Millsap & Meredith, in press) are that there may be significant problems in utilizing the MH procedure in data sets following multi-parameter ICC models. The variables selected for examination in this study were designed to verify results of this theoretical work. As an approximation procedure, MH is potentially sensitive to a number of influences which impact on its effectiveness. Certain of these influences are well understood and predictable such as the inclusion/exclusion of the studied item, number of test items and number of examinees. The number of subjects in focal and reference groups were equal in this study to control for such known influences, but typically there are fewer examinees in the focal group. Thus, certain real data sets may not provide stable statistical referents. The effects of sample size on the variances of α_{mh} are also well understood, as is the performance of the ETS DIF classification scheme in controlling for practical significance. However, there has been no work on the effects of ICC parameterizations of studied and criterion items or Θ distributions of focal and reference groups on the variances or the ETS DIF classifications. The variances influence the consistency of ETS DIF classifications through the ETS significance test of Δ_{mh}. Therefore, these variables were selected for inclusion in this study.

The results of Phase I clearly indicate that Θ_{dis}, the interaction of the A and C parameters and the third order

interaction between θ_{dis} and the A and C parameters significantly influence the α, Δ_{mh} and the MH_{chisq} type I error rate, at the .05 level. This effectively means that the procedure is detecting DIF where none is modelled, because of certain ICC characteristics, θ distributions, and the interaction between these factors. However, DIF should not be detected where none exists and the procedure is supposed to control for ability. Therefore this study calls into question the overall effectiveness of the MH procedure in the type I case under conditions approaching those modelled.

In the type II cases the results are considered within the context of a set reference group ICC (1, 0, .2) and variation of the focal group ICC as per the design. The finding that the A, B and C ICC parameter main effects are significant arises for two reasons. For the non-uniform DIF cases, they interact with the reference group parameter values to produce more or less a degree of non-uniform DIF. This in turn results in higher type II error rates. It is important to note that the inability of the MH procedure to detect items having non-uniform DIF is not an all or nothing issue. The extent of non-uniform DIF and the locations of the θ distributions for the reference and focal groups relative to the θ scale are important influences on the type II error rate under non-uniform DIF conditions.

For the uniform DIF type II cases A, B and C ICC parameter factors affect the relationship between the location

and form of DIF relative to the locations of the Θ distributions. This in turn impacts on the various MH statistics under study.

For the Phase II Rasch case results we find that the procedure is behaving as expected in that the levels of α and Δ_{mh}, a function of α, remain the same regardless of the nature of the conditions modelled. The level of α and Δ_{mh} reflect the modelled DIF condition, consistently detecting a constant amount of bias. However, this is not the only important issue regarding the Rasch case. This study shows that the level of the MH chi square is dramatically affected by the increasingly discrepant relationship, over the cases, between the b parameter (difficulty) of the studied item for focal and reference groups and 1) range of b parameters of items in M and 2) Θ distributions of the reference and focal groups. These disturbances occur within the context of detecting a constant amount of DIF and are reflected in the type II error rate. The discrepant conditions simulated in this phase result in increasingly lower probabilities of most examinees passing the studied item and the items in M, the matching criteria. This results in low item and test information which affects the MH chi square and the type II error rates.

The results of the third phase of the study are evaluated by considering the performance of $Var_{ht}\ln(\alpha)$ relative to $Var_{emp}\ln(\alpha)$, under the type I and type II conditions, respectively. The influences on $Var_{emp}\ln(\alpha)$ are considered the

real effects of the simulated conditions. Under the type I conditions $Var_{emp}ln(\alpha)$ is significantly affected only by the A, C and θ_{dis} factors and the C by θ_{dis} interaction. $Var_{ht}ln(\alpha)$, however is significantly affected by the A, B, C and θ_{dis} main effects and the A by C, A by θ_{dis}, B by C, B by θ_{dis}, C by θ_{dis} and A by C by θ_{dis} interactions. Therefore $Var_{ht}ln(\alpha)$ is significantly affected above and beyond the real influences on the actual variance by the B main effect and the A by C, A by θ_{dis} B by C, B by θ_{dis} and A by C by θ_{dis} interactions. This would appear to be problematic as the estimated variance, $Var_{ht}ln(\alpha)$, has several additional and statistically significant influences in the type I case leading to potential inaccuracies. Related measures such as the ETS DIF classification scheme could also then be adversely affected.

In the type II error rate case the performance of $Var_{ht}ln(\alpha)$ appears to be more promising. $Var_{emp}ln(\alpha)$ is significantly influenced by the A, C and θ_{dis} main effects and the A by C, B by C, C by θ_{dis} and A by B by C interactions. $Var_{ht}ln(\alpha)$ is influenced by these factors along with one more, the B main effect. The additional influences of this factor may be important, but the overall disparity of significant main effects and interactions is not as large as in the type I case. The appropriate interpretation appears to be that $Var_{ht}ln(\alpha)$ is more similarly influenced by the design features of the data sets, compared to $Var_{emp}ln(\alpha)$, in type II cases than the type I cases.

The consistency of the ETS DIF classification scheme is related to the magnitude of $Var_{emp}ln(\alpha)$ and $Var_{ht}ln(\alpha)$. The pattern of results for type I error rate cases MH38A and MH39A suggest that high levels of variance effect the consistency of classification of misclassified items. Under certain conditions this could lead to high levels of confidence on the part of test evaluators that items are problematic (frequently misclassified as "C") when in fact no DIF would be present.

The results of phase IV show that the parameterizations of the items in the matching criteria, M, have important effects on the capability of the procedure to detect modelled DIF and no DIF conditions. For the type I and type II error rate cases, the θ_{dis} factor is a consistent overall influence on the MH statistics as would be expected from the results of phase I. The simulations performed for phase IV are an entirely separate and independent set and therefore the presence of a significant θ_{dis} main effect result provides additional and weighty evidence that the MH procedure is adversely affected by differences in the ability distributions in both the type I and type II cases.

The finding most specific to this phase is the significant θ_{dis} by Eta interaction in both the type I and type II cases. Eta^2 gives an indication of the utility of scores on the criterion items as a proxy for latent θ ability distributions for reference and focal groups. The ICC parameterizations of the criterion items therefore have

important influences on the MH statistics through their effects on eta and interaction with θ_{dis}. This represents another important problem in the application of the procedure since the assumption seems to be made that the score on criterion items approximates θ well. The level of eta is not known in MH DIF applications. Since there is an important interaction of eta with θ_{dis} and focal and reference group θ distributions are not typically equivalent there is potential for serious inaccuracies in the application of the procedure.

VI. Summary

The Mantel Haenszel procedure is intended to approximate the theoretically preferred IRT definition of DIF (Lord, 1980). IRT DIF models condition on ability and test the equivalence of item parameters between the focal and reference groups. The study reported here suggests that there are significant problems in utilizing the Mantel Haenszel procedure to recover IRT DIF.

The item parameterizations of the studied items and the items in the matching criteria have clear effects on the accuracy of the procedure in both the type I and type II cases. In addition, ability discrepancies between the focal and reference groups affect the procedure through item and test information. The practical implication of these findings is that caution should be used in interpreting the presence or absence of DIF in items in data sets which may follow two and three parameter IRT models. This is especially important when significant ability as total score discrepancies are apparent.

When the data follow the Rasch model, α_{mh} behaves in the expected manner in that it is not affected by ability discrepancies. However, this is not true of the MH chi square significance test. The results show a clear trend toward smaller chi square values for data sets with ability discrepancies and extreme difficulty values of the items in the matching criteria. Prudence should be used in interpreting MH_{chisq} results under such circumstances. The p

values of items across tests expected to follow the Rasch model may give an indication of conditions warranting caution.

The estimated variance (Phillips & Holland, 1987) is influenced by item parameterizations and ability discrepancies beyond the extent to which the empirical variance is affected by these factors. This is especially apparent in the type I case. Therefore researchers should be mindful of possible inaccuracy of $Var_{ht}\ln(\theta)$ in data sets possibly following two and three parameter IRT models. The level of the variances also affects the stability of ETS DIF classifications. Therefore, researchers should be alert to possible instability under high variance conditions.

The reliability of total score given known ability estimates, eta, also impacts on the accuracy of the procedure. The interaction of eta with ability discrepancies is significant in both the type I and type II cases. In terms of practical implications, researchers should be aware that total score may not serve as an adequate proxy for the latent ability distributions of the focal and reference groups.

The study presented here is intended to compliment the work of Donaghue et al (in press). Further simulation research should focus on the interactions of factors examined in their work and the study presented here. For example, the interactions between the ICC parameterizations, ability discrepancies, number of test items, and absolute and relative sizes of focal and reference groups could be examined.

APPENDIX A

The RANGEN Program

Rangen generated pseudo-random item data to fit a 1, 2 or 3 parameter logistic IRT model. The number of items and examinees and item paramter values (a, b, c) are specified, along with a seed value which begins data generation. Person ability (θ) parameter values are randomly generated from normal prior distributions with specifiable mean and variance. Output includes the examinee by item matirix, item p values and IMSL output on the regression of person ability parameters on total score.

Data for the focal and reference groups are generated, sample sizes may differ between the two groups but the number of items must be the same. The Mantel Haenszel chi-square, alpha, delta and estimated (Phillips and Holland, 1986) and empirical variances of alpha are output for the first item. In addition the number of iterations categorized under each of the ETS DIF classifications is output. The IMSL subroutine outputs eta values for the reference and focal group data sets.

To generate simulated examinee item scores the probability of an examinee answering the item correctly was assumed to be given by

$$P_i(\theta) = c_i + (1-c_i)\,[1+e^{-Da_i(\theta-b_i)}]^{-1}.$$

Since θ and the item parameter values are known $P_i(\theta)$ is

determined. The item score is then generated when the program selects a random number from the uniform distribution [0,1]. If the random number is less than or equal to $P_i(\theta)$, which happens $P_i(\theta)$ of the time, the examinee is scored 1, otherwise the examinee is scored 0. The process is repeated for each item and examinee making available the data sets from which RANGEN calculates the various statistics.

REFERENCES

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

Burrill, L.E. (1982). Comparitive studies of item bias methods. In R.A. Berk (Ed.), Handbook of Methods for Detecting Test Bias, Baltimore, MD: The Johns Hopkins University.

Cleary, T.A. & Hilton, T.L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Camilli, G & Smith, J.K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. Journal of Educational Statistics, 15(1), 53-67.

Devine, P.J. & Raju, N.S. (1982). Extent of overlap among four item bias methods, Educational and Psychological Measurement, 42, 1049-1066.

Donlon, T.F. & Angoff, W.H. (1971). The Scholastic Aptitude Test. In W.H. Angoff (Ed.), The College Board Admissions Testing Program New York: College Entrance Examination Board.

Donaghue, J., Holland, P.W. & Thayer, D.T. (in press). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential Item Functioning: Theory & Practice. Hillsdale, NJ: Erlbaum

Green, D.R. & Draper, J.F. (1972, September). Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu,

Hambleton, R.K. & Rogers, H.J., (1988). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334

Harvey, A. (1990, April). The stability of the Mantel-Haenszel D-DIF statistic across populations differing in ability. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

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

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). Item Response Theory: Application to Psychological Measurement, Homewood, IL: Dow Jones-Irwin.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk (Ed.), Handbook of Methods for Detecting Item Bias Baltimore, MD: The Johns Hopkins University.

Ironson, G.H. & Subkoviak, M. (1979). A comparison of several methods assessing item bias. Journal of Educational Measurement, 16, 209-225.

Linacre, J.M. & Wright, B.D. (1987). Item Bias: Mantel-Haenszel and the Rasch Model, (Memorandum No.39), MESA Psychometric Laboratory, Department of Education, University of Chicago.

Linn, R.L. & Harnish, 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. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems, Hillsdale, NJ: Erlbaum.

Lyman, H.B. (1986). Test Scores and What They Mean. (4th Ed.). New Jersey, Prentice Hall.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-740.

Marascuilo, L.A. & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on X^2 statistics. Journal of Educational Measurement, 18, 229-248.

Meredith, W. & Millsap, R.E. (in press). On the misuse of manifest variables in the detection of measurement bias. Psychometrika.

Millsap, R.E. & Meredith, W. (in press). Inferential conditions in the statistical detection of measurement bias. Applied Psychological Measurement.

Mislevy, R.J. & Bock, R.D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items. Chicago: International Educational Services.

Phillips, A. & Holland, P.W. (1987). Estimation of the variance of the Mantel-Haenszel log-odds ratio estimate. Biometrics, 43, 425-431

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

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

Rudner, L.M. (1977, April). An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

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

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

Schulz, E.M., Perlman, C.P., Rice, W.K. & Wright B.D. (1989, April) Empirical comparison of Rasch and Mantel Haenszel procedures. Annual Meeting of the American Educational Research Association, Boston, MA

Sharif by Salahuddin vs. New York State Education Department. 709 F.Supp 345 (S.D.N.Y., 1989).

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

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

Spray, J.A. (1989). Performance of three conditional DIF statistics in detecting differential item functioning on simulated tests. (Research Report Series 89-7). American College Testing Program.

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

Swaminathan, H. & Gifford, J.A. (1979). Estimation of parameters in the three parameter latent trait model. (Report No.90) Laboratory for Psychometric and Evaluation Research, Amherst, MA: University of Massachusetts, School of Education.

Thissen, D. & Wainer, H. (1990). Confidence envelopes for item response theory. Journal of Educational Statistics, 15(2), 113-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.) Test Validity, Hillsdale, NJ: Erlbaum.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). Logist User's Guide, Princeton, NJ: Educational Testing Service.

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

Wright, B. & Stone, M. (1979). Best Test Design, Chicago, IL: MESA.

Wright, B. & Stone, M. (1988). Identification of item bias using Rasch measurement. (Research Memorandum No. 55). Chicago: University of Chicago, Statistical Laboratory, Department of Education, MESA.

Zwick, R. (1990). When do Item Response Function and Mantel-Haenszel Definitions of Differential Item Functioning Coincide? Journal of Educational Statistics, 15, 185-197.