



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file - Votre référence

Our file - Notre référence

NOTICE

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

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

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

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

AVIS

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

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

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

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

**A PRESENTATION AND COMPARISON OF SOME NEW
STATISTICAL TECHNIQUES IN THE ANALYSIS OF POLYTOMOUS
DIFFERENTIAL ITEM FUNCTIONING: A MONTE CARLO INVESTIGATION**

By

Louise C. Mâsse

Dissertation presented to the School of Graduate
Studies and Research as partial fulfillment
of the Ph.D. degree in Education

Faculty of Education, University of Ottawa
Ottawa, Canada, 1994

© Louise C. Mâsse, Ottawa, Canada, 1994



National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services Branch

Direction des acquisitions et
des services bibliographiques

355 Wellington Street
Ottawa, Ontario
K1A 0N4

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

THE AUTHOR HAS GRANTED AN
IRREVOCABLE NON-EXCLUSIVE
LICENCE ALLOWING THE NATIONAL
LIBRARY OF CANADA TO
REPRODUCE, LOAN, DISTRIBUTE OR
SELL COPIES OF HIS/HER THESIS BY
ANY MEANS AND IN ANY FORM OR
FORMAT, MAKING THIS THESIS
AVAILABLE TO INTERESTED
PERSONS.

L'AUTEUR A ACCORDE UNE LICENCE
IRREVOCABLE ET NON EXCLUSIVE
PERMETTANT A LA BIBLIOTHEQUE
NATIONALE DU CANADA DE
REPRODUIRE, PRETER, DISTRIBUER
OU VENDRE DES COPIES DE SA
THESE DE QUELQUE MANIERE ET
SOUS QUELQUE FORME QUE CE SOIT
POUR METTRE DES EXEMPLAIRES DE
CETTE THESE A LA DISPOSITION DES
PERSONNE INTERESSEES.

THE AUTHOR RETAINS OWNERSHIP
OF THE COPYRIGHT IN HIS/HER
THESIS. NEITHER THE THESIS NOR
SUBSTANTIAL EXTRACTS FROM IT
MAY BE PRINTED OR OTHERWISE
REPRODUCED WITHOUT HIS/HER
PERMISSION.

L'AUTEUR CONSERVE LA PROPRIETE
DU DROIT D'AUTEUR QUI PROTEGE
SA THESE. NI LA THESE NI DES
EXTRAITS SUBSTANTIELS DE CELLE-
CI NE DOIVENT ETRE IMPRIMES OU
AUTREMENT REPRODUITS SANS SON
AUTORISATION.

ISBN 0-315-95958-4

Canada



UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Acknowledgments

I would like to thank my advisor, Dr. Marc E. Gessaroli, for his support, patience, and endless encouragement in the completion of this work. Among other things, his availability and suggestions were greatly appreciated. I would like to take this opportunity to thank him for making the completion of this work enjoyable and fun. I feel very privileged to have worked with him and I have learned a great deal from him.

I would also like to thank Dr. Marvin Boss and Dr. Bruno Zumbo for their useful recommendations and criticisms. Their comments were very useful and have helped to make this work more valuable.

I am also grateful to Daniel Charbonneau from the University of Montreal for his valuable computer help. Without his telnet expertise I would not have been able to collect my data from Montreal and would probably still be travelling on the 417. I would also like to thank the CRT of the University of Montreal for allowing me to use their computer facilities.

A special thank you goes to my parents for their never ending encouragement and positive support. Their own desire to learn and their attitude toward education have always motivated me and been my driving force. I would also like to thank them for helping me on my many moves over the years. Mom, I was very fortunate that you always seemed to have the right connections when I had computer problems.

Finally, a very special thank you goes to Jeff whose patience and motivation have made my life a breeze. Yes! I didn't forget! I owe you big time for all the proof reading you did for me. I definitely would not have your patience.

Table of Contents

	Page
ABSTRACT	viii
CHAPTER 1 : INTRODUCTION	1
Statement of the problem	1
Purposes of the study	5
Organization of this work	5
CHAPTER 2 : REVIEW OF LITERATURE	7
Test bias versus item bias	7
Problems in the identification of item bias	8
Perspectives in the identification of Item	
Response Differences (IRD)	10
• Defining uniform and nonuniform PDIF	13
The need to study PDIF	15
Review of existing PDIF procedures	19
• Log-Linear approach	19
• Veale and Foreman index	20
• Welch and Hoover procedures: The HW1, HW3,	
and the Mantel-Haenzsel	21
• Standardization approach	27
• Logistic regression	27
• IRT models	30
• Summary of the approaches presently used	
to assess PDIF	32
CHAPTER 3 : DESCRIPTION OF THE PDIF APPROACHES STUDIED	34
Formulation of a general example	35
• The MAN(C)OVA approaches	37
• MANCOVA (MCO) approach	38
• MANOVA (MCA) approach	45
• Empirical support for the MCA and MCC	
approaches	46
• The Dual Scaling (DS) approach	49
• The Polytomous Logistic Regression (PLR)	
approach	51
• Empirical support for the Polytomous	
Logistic Regression approach	56
• The log-linear (LOG) approach	58
Relationship among the PDIF approaches studied	60
CHAPTER 4 : METHODS	64
The PDIF approaches	64
• Conditioning strategies	65
Simulation conditions	66
• Factors held constant	66
• Number of groups	66
• Ability distribution	67

	Page
• Number of items exhibiting PDIF	67
• Replications	67
• Factors manipulated	68
• Test length	68
• Sample size	68
• Item difficulty	69
• Amount and location of PDIF	69
• Final design	72
Generation of the item responses	72
• The two parameter polytomous logistic regression model	73
• Item parameters used to generate the data	75
• Generating the item responses	77
Computer programs	79
• Item response program	79
• Computing the DS approach	79
• Computing the LOG approach	80
• Computing the MCA and MCO approaches	81
• Computing the PLR approach	82
Uniform and nonuniform test of hypothesis	83
• Logit differences	83
• Differences in proportions	83
• Type of PDIF simulated	84
• PDIF 1 conditions	85
• PDIF 2 conditions	90
Data analysis	94
Research questions	95
Limitations of the study	95
 CHAPTER 5 : RESULTS	 99
Data generation verification	101
• Verification of the IMSL subroutines	102
• GGNML subroutine	103
• GGUBS subroutine	104
• Verification of the data generation program	105
• Total test score descriptive results	107
• Ability group distributions	109
• Input and output parameters	110
Test of uniform PDIF	113
• NO PDIF conditions	113
• PDIF conditions	115
Test of nonuniform PDIF	118
• NO PDIF conditions	119
• PDIF conditions	123
 CHAPTER 6 : DISCUSSION	 132
Test of uniform PDIF	133
• NO PDIF conditions	133
• PDIF conditions	134
• Item difficulty	134

• Location of PDIF	136
• Sample size	137
• Test length	137
Test of nonuniform PDIF	138
• LOG approach	139
• MCA approach	141
• MCO approach	142
• PLR approach	143
Contrasting the approaches	144
Comparing the results with other studies	149
CHAPTER 7 : CONCLUSIONS	152
Uniform tests of hypothesis	152
Nonuniform tests of hypothesis	153
Overall conclusions	154
Suggestions for future research	156
REFERENCES	159
APPENDIX A	169
APPENDIX B	178
APPENDIX C	183
APPENDIX D	186

List of Tables

NO	NAME	Page
1	Illustration of a contingency table at ability level i , where $i=1,2,\dots,k$	24
2	A response-pattern matrix for a multiple-choice test	35
3	Example of the J matrix for item i	43
4	Proportion (difficulty parameter) and biserial correlation (discrimination parameter) for the 0.70 PDIF conditions	70
5	Proportion (difficulty parameter) and biserial correlation (discrimination parameter) for the 0.40 PDIF conditions	70
6	Example of an indicator matrix for item i	81
7	Summary of GGNML verification	103
8	Summary of GGUBS verification	104
9	Descriptive statistics for the basis, reference, and focal groups	107
10	Ability group distributions for 100 replications	109
11	Summary of the regression analyses among the reference (R), focal (F), and basis (B) groups	110
12	Uniform tests of hypothesis for the NO PDIF conditions	114
13	DS average number of Type I errors for the test length by item difficulty interaction	115
14	Numbers of rejections for the 0.70 PDIF conditions	117
15	Numbers of rejections for the 0.40 PDIF conditions	117
16	Nonuniform tests of hypothesis for the NO PDIF conditions	120

NO	NAME	Page
17	PLR average number of Type I errors for the test length by item difficulty interaction . . .	121
18	MCA average number of Type I errors for the test length by sample size interaction	122
19	Nonuniform tests of hypothesis for the 0.70 PDIF conditions	124
20	Nonuniform tests of hypothesis for the 0.40 PDIF conditions	124
21	PLR average number of Type I errors for the test length by PDIF interaction	126
22	ANR results for the 0.70 PDIF conditions	129
23	ANR results for the 0.40 PDIF conditions	130
24	Differences in proportions and b parameter introduced in the focal group	135
25	Example of expected proportions for an item . .	140

List of Figures

NO	NAME	Page
1	Contrast 1 characteristic curves for the PDIF 1 condition (proportion)	86
2	Contrast 2 characteristic curves for the PDIF 1 condition (proportion)	87
3	Contrast 3 characteristic curves for the PDIF 1 condition (proportion)	87
4	Contrast 1 characteristic curves for the PDIF 1 condition (logit)	88
5	Contrast 2 characteristic curves for the PDIF 1 condition (logit)	89
6	Contrast 3 characteristic curves for the PDIF 1 condition (logit)	89
7	Contrast 1 characteristic curves for the PDIF 2 condition (proportion)	91
8	Contrast 2 characteristic curves for the PDIF 2 condition (proportion)	91
9	Contrast 3 characteristic curves for the PDIF 2 condition (proportion)	92
10	Contrast 1 characteristic curves for the PDIF 2 condition (logit)	92
11	Contrast 2 characteristic curves for the PDIF 2 condition (logit)	93
12	Contrast 3 characteristic curves for the PDIF 2 condition (logit)	93
13	GGNML standard deviates distribution	103
14	Distribution of the random numbers	105

Abstract

Many methods have been suggested to detect items that perform differently across subgroups of examinees. These methods fall under the general heading of Item Response Differences (IRD). However, most of the IRD detection methods require item responses that have been dichotomously coded (e.g. correct/incorrect). In the case of multiple-choice items there may be some loss of information when the responses are dichotomously coded because the information found in the distractors (incorrect options) is not utilized. It, therefore, appears worthwhile to use methods of analysis that use information in all of the response options in the detection of IRD. Because few approaches have been suggested for analyzing IRD in this manner there is a need to develop and investigate methods which can assess the IRD found in all the options of an item. In this study, such an investigation was referred to as Polytomous Differential Item Functioning (PDIF). The purpose of this study was to present and investigate the performance of four new approaches in the assessment of PDIF.

The four approaches are a MANOVA (MCO) and a MANCOVA (MCA) approach applied to categorical dependent variables, a Polytomous Logistic Regression (PLR) approach, and an ANOVA analysis based on the item responses quantified by Dual Scaling (DS). In this study the effectiveness of

these approaches (MCA, MCO, PLR, and DS) as well as the Log-Linear (LOG) approach of Mellenbergh (1982) were assessed under various conditions of test length, sample size, item difficulty, and the amount and location of PDIF.

A two-parameter polytomous logistic regression model was used to generate the data. In this study, only uniform PDIF was introduced in the alternatives of the item.

The type of PDIF simulated (e.g. uniform) in this study did not allow for a direct comparison of the nonuniform test of hypothesis between the Logistic (LOG and PLR) approaches and the MAN(C)OVA (MCA and MCO) approaches because the Logistic approaches test for a difference in logits while the MAN(C)OVA approaches test for a difference in proportions. It was shown in this study that varying the probability of choosing the alternatives resulted in uniform logit differences which did not only translate into uniform differences in proportions but also translated into nonuniform differences in proportions. These differences affected the interpretation of the PDIF results because the test of nonuniform PDIF for the Logistic procedures corresponded to a valid test of the null hypothesis while the MAN(C)OVA results for nonuniform PDIF had to be adjusted in order to yield a test which approximated a true test of the null hypothesis.

The results of this study indicated that for the test of uniform PDIF the LOG, MCA, MCO, and PLR approaches had adequate Type I error rates as well as being equally powerful. The results of the DS approach also indicated good Type I error control for the test of uniform PDIF. However, its power of detection was much lower than for the other approaches.

The power of detection for the test of uniform PDIF was affected by the difficulty of the item, the size of the samples, and the location of PDIF for the LOG, MCA, MCO, and PLR approaches. The power of detection for the DS approach was affected by the size of the samples, the location of PDIF, and by the length of the test. The lack of power associated with the DS approach was believed to be related to the low number of items exhibiting PDIF (i.e. 1 item).

The Type I error rates for the test of nonuniform PDIF with the NO PDIF and PDIF conditions were mainly within the expected nominal range for the PLR approach, while the LOG approach had larger than expected Type I errors in some conditions. For the MCA and MCO approaches, conclusions were only possible for the nonuniform NO PDIF conditions because the nonuniform PDIF conditions did not correspond to a true test of the null hypothesis. Of these two approaches the MCA was the only one which was found to maintain adequate Type I error rates. The effects of the

independent variables on these Type I error rates were not conclusive for most of these approaches. There were several reasons for this: 1) The effects found for the nonuniform test of hypothesis with the PDIF and NO PDIF conditions differed. Because no nonuniform logit differences were introduced in the alternatives these tests should correspond to a test of the null hypothesis for both the PDIF and NO PDIF conditions. Therefore, the effect of the independent variables for the nonuniform test of hypothesis should be consistent for the PDIF and NO PDIF conditions. This was not observed in most cases. 2) It is yet uncertain as to how much weight should be attributed to significant effects which are computed on rejection rates that are clearly within the expected nominal range. 3) The MCA and MCO approaches yield only approximate nonuniform tests of the null hypothesis for the PDIF conditions, and therefore, these results should be replicated with a true test of the null hypothesis before any strong conclusions regarding the effects of the independent variables are made. Nonetheless, the Type I error rates for the MCO seem to be clearly affected by the difficulty of the item.

In conclusion, the results of this study lend some optimism to the employment of the MCA and PLR approaches. Because the conclusions of this study are limited to the scope of this study it is suggested that more research be

conducted in this area in order to ascertain the conditions in which these approaches may be effectively used.

CHAPTER I Introduction

Statement of the problem

Everyday, tests are used to make important decisions. Sometimes, this is done without knowing if the tests display different characteristics depending upon gender, cultural, or social background. Because tests are often used for selection purposes, concern about the differences in test performance among certain groups appears to be well justified (Angoff, 1993). Also, for obvious legal reasons, it is important to develop tests that do not offer unwanted selection advantages to certain groups. For these reasons, assessing bias has long been the concern of many researchers over the years. There are many ways into which bias in a test can be assessed. This study is mainly concerned with the assessment of bias at the item level.

Statistical methods are usually employed to assess item bias. However, item bias cannot be fully defined in statistical terms. Statistical techniques may suggest that item bias is present but the item must be further examined before item bias is said to exist. To acknowledge that statistical techniques only assess items which are potentially biased, researchers have referred to these analyses as Differential Item Functioning (DIF) (Green, 1991; Hills, 1989; Holland & Thayer, 1986; Shepard, 1982) or Differential Item Performance (DIP) (Welch, Ackerman, &

Doolittle, 1987; Wright, 1986). These terms over the years have been associated with the analysis of only the correct answers. Although, this has been the method of choice over the years other methods of analysis may be employed to identify items which are potentially biased. Some researchers have suggested the analysis of the distractors or the incorrect alternatives (Green, Crone, & Folk, 1989; Veale & Foreman, 1983) and other have suggested the analysis of all the alternatives (Dorans, Schmitt, & Bleistein, 1992) as a way of assessing items which may be potentially biased. These analyses have been referred to as Differential Distractor Functioning (DDF) and Differential Alternative Functioning (DAF), respectively. All these methods for assessing items which are potentially biased may be classified as different methods of assessing Item Response Differences (IRD). This term, namely IRD, will be used in this study to refer to items which have been identified as being potentially biased. This term is suggested because it does not make any reference as to the method of analysis used.

Studies conducted in the area of IRD have routinely applied a binary model to data from a multiple-choice test (Thissen & Steinberg, 1984). In doing this, much of the information available in the distractors is lost (Green, Crone, & Folk, 1989; Thissen, 1976; Thissen & Steinberg, 1984; Thissen, Steinberg, & Fitzpatrick, 1989; Veale &

Foreman, 1983). Evidence that the distractors may provide some information is found in the work of Bock (1972), Simpson (1983), Thissen (1976) and Thissen and Steinberg (1984). Knowing that some information exists in the distractors, some researchers (e.g. Green, Crone, & Folk, 1989; Veale & Foreman, 1983) have used the distractors in the method to analyze IRD. However, this type of analysis ignores the differences which are found in the correct options. This may be resolved by conducting an analysis which includes all the alternatives. Actually, including all the alternatives in the analysis can provide more and better information concerning IRD as compared to analyzing only the responses to the correct answer or to the incorrect alternatives. This approach to the analysis of IRD is the focus of this study. In this study, the IRD analysis of all the alternatives is referred to as Polytomous Differential Item Functioning (PDIF).

Statistical methods that have been proposed for the analysis of PDIF may be divided into three categories. The statistical methods in the first category are those which were developed for the analysis of IRD that are found in essay items (Welch & Hoover, 1991; Wilson, Spray, & Miller, 1993). Most of these methods require that the data be measured at least on an ordinal scale. The statistical methods in the second category were developed to study DIF. These methods essentially suggest using dichotomously

item-scored DIF methods to analyze PDIF (Dorans, Schmitt, & Bleistein, 1992; Wilson, Spray, & Miller, 1993). The statistical PDIF methods within the last category are those that have been suggested for the analysis of multiple-choice items (i.e. nominal data). This last category of analyses is the one of interest in this study. Although, the statistical methods in the second category may be used to analyze PDIF found in multiple-choice items, they are not considered in this study because analyzing PDIF in this fashion offers questionable control of the Type I error rates.

In the literature, few statistical procedures have been suggested for the analysis of PDIF found in multiple-choice items. Actually, the Log-Linear approach proposed by Mellenbergh (1982) appears to be the only method which has potential to analyze PDIF which is found in multiple-choice items. Therefore, further research in this area should focus on developing statistics which can be used for the analysis of PDIF when the item responses are nominal. Also, because the utility of the Log-Linear approach presented by Mellenbergh (1982) has not yet been fully investigated there appears to be a need to further examine this procedure.

Purposes of the study

The purposes of this study were twofold. The first purpose was to present some new statistical procedures that might be used for the analysis of PDIF found in multiple-choice items. These new approaches are, specifically, the Polytomous Logistic Regression approach (PLR), the Dual Scaling approach (DS), the MANOVA (MCA) and the MANCOVA (MCO) approaches applied to categorical dependent variables.

The second purpose of this study was to investigate the effectiveness of the Dual Scaling, MANOVA, MANCOVA, and Polytomous Logistic Regression approaches in their assessment of PDIF under various conditions of test length, sample size, item difficulty, and the amount and location of PDIF. The effectiveness of these new approaches were also to be compared with the performance of the Log-Linear (LOG) method proposed by Mellenbergh (1982).

Organization of this work

In Chapter 2 the various perspectives which have been suggested to study IRD are presented. The terminology associated with these perspectives is presented, as well as some elaboration of the perspective under which IRD was investigated in this study. Next, the need to study IRD as suggested in this study is provided. This is followed by a

review of the techniques that have been previously proposed for the analysis of PDIF.

In Chapter 3 the rationale underlying the statistical approaches studied for the analysis of PDIF are presented. The theoretical foundation underlying these approaches is presented as well as some empirical evidence for the use of these approaches in the analysis of nominal item responses. This chapter concludes with a comparison of the approaches.

In Chapter 4, the methodology used to assess the effectiveness of the approaches is presented. Included are the conditions under which the statistics were investigated. Also, in this chapter the limitations associated with this study are discussed.

The results of this investigation are presented in Chapter 5 in addition to some evidence regarding the adequacy of the data generation procedures is provided.

The results are then discussed in Chapter 6. In this chapter the approaches are compared in their effectiveness to assess PDIF. This chapter concludes by comparing the results of this study with findings from other studies.

Finally, the conclusions which can be drawn from this study are presented in Chapter 7. Suggestions for future research are also found in this chapter.

CHAPTER 2

Review of Literature

The purposes of this chapter are first to present how "bias" found in a multiple-choice test have been studied, as well as identifying how it was studied in this investigation. This is followed by a clarification and presentation of the terminology that has been employed for the analysis of item response differences. Then, the need to assess item response differences as was done in this study is discussed. The chapter concludes with a review of previously proposed techniques for the analysis of PDIF.

Test bias versus item bias

Studies conducted in the area of bias have analyzed bias in two different ways. Test bias consists of focusing on the bias found in the whole test. Test bias studies (e.g. Cole, 1973; Darlington, 1971; Linn, 1971, 1984; Shealy & Stout, 1993; Thorndike, 1971) consist of investigating group differences by focusing on the differences found in the total test scores. These studies can be, generally, described as investigating whether the test scores for different groups of equal ability have the same meaning (Green, 1991; Linn, 1971; Shepard, 1982). Item bias consists of focusing on the individual items themselves as a method for locating the bias which is

found in a test. The purpose of item bias studies, generally, is to identify items that function differently among subgroups of examinees of comparable ability (Green, 1991; Hills, 1989; Shepard, 1982; Shepard, Camilli, & William, 1984). This type of bias is the focus of this investigation.

Problems in the identification of item bias

Item bias has been defined by several authors (Ironson, 1982; Linn & Drasgow, 1987; Linn, Levine, Hasting, & Wardrop, 1981; Shepard, 1982; Shepard, Camilli, & Averill, 1981) in terms of the probability of answering an item correctly. More specifically, an item is said to be biased if examinees of equal ability, from different groups, do not have equal probabilities of correctly answering the item. Psychometricians have developed several statistical methods in an attempt to identify biased items in a test. These statistical methods have the potential to identify item bias only if the ability of the subjects has been considered. Otherwise one may identify an item as being potentially biased when in fact only differences in group abilities are present (i.e. impact). However, even when the ability of the subjects has been accounted for, these statistics do not necessarily ascertain that item bias has been identified. These statistical methods only identify the items which are

functioning differently for the population subgroups under study. This may, in turn, suggest that item bias is in fact present in the sense that the item is judged to offer unfair advantages for one of the subgroups. However, the item must be further analyzed substantively in order to judge if in fact bias is found in the item. The items may have been found to function differently against one subgroup of examinees because they called on important educational outcomes which are unequally known to the students or for some other unknown reasons which are not related to bias. However, the word bias itself contains the sense of an evaluation which goes beyond the differences in performance observed among the groups (Angoff, 1993; Burrill, 1982). The word bias mainly implies that one group is being disadvantaged because of gender, social, or perhaps, cultural differences. Bias is, however, not implied when the difference observed is related to other external criteria such as curriculum differences.

Therefore, it is quite obvious that the term item bias cannot be used to define statistical differences (Angoff, 1993; Burrill, 1982; Holland & Thayer, 1986; Wright, 1986). Some suggestions were made to use a term other than item bias to refer to statistical observations which would not imply any judgemental or interpretive meaning (Angoff, 1983; Burrill, 1982). For this purpose

the expression **Item Response Differences (IRD)** is used in this study to refer to items which have been identified as functioning differently for examinees of comparable abilities, but belonging to different subgroups. This expression is used because it is more general than the usual expressions (i.e. Differential Item Functioning or Differential Item Performance). These terms, as discussed in the next section, have been associated over the years with specific methods of analysis.

Perspectives in the identification of Item Response Differences (IRD)

Statistical methodologies developed to study IRD can be divided into four different perspectives. These perspectives as well as the terminology associated with these are presented in this section.

The first and most common method of studying IRD consists of focusing on the pattern of only the correct options. This type of IRD analysis has been referred to as **Differential Item Functioning (DIF)** (Green, 1991; Hills, 1989; Holland & Thayer, 1986; Shepard, 1982) or **Differential Item Performance (DIP)** (Welch, Ackerman, & Doolittle, 1987; Wright, 1986). An item is said to exhibit DIF if examinees of comparable abilities, belonging to different subgroups, do not have the same probability of getting the item correct (Holland & Thayer, 1986; Shepard,

1982). The definition of DIF differs from IRD only in that DIF specifies how the differential item functioning has been assessed (i.e. in the correct answer). The definition of IRD is more global in that it does not specify how the differences were assessed. It should be mentioned that originally the definition of DIF or DIP was probably not intended to refer only to the differences observed in the correct options. However, because over the years most methods used to assess DIF or DIP mainly examined the pattern of the correct options, the terms DIF or DIP tend to be associated with only the analysis of the correct options.

Another way of analyzing IRD contained in the items is to focus on the analysis of the incorrect alternatives (or distractors) in a multiple-choice item (Green, 1991; Green, Crone, & Folk, 1989; Veale & Foreman, 1983). Green, Crone, and Folk (1989) defined this type of analysis as **Differential Distractor Functioning (DDF)**. The rationale underlying DDF is the same as DIF. DDF consists of assessing if the groups under study are attracted differently to the distractors of an item, when the ability of the individuals is controlled. Some authors (Green, 1991; Green, Crone, & Folk, 1989; Veale & Foreman, 1983) recommended this approach to study IRD, while others use this approach to explain the reason or reasons

underlying DIF (Dorans & Holland, 1993; Dorans, Schmitt, Bleistein, 1992).

Recently, more statistical methods have been developed to analyze IRD found in all the alternatives of a multiple-choice item. This approach has been referred by some authors as **Differential Alternative Functioning (DAF)** (Thissen, Steinberg, & Wainer, 1993), or **Comprehensive Differential Item Functioning (CDIF)** (Dorans, Schmitt, & Bleistein, 1992). The presentation of new procedures which can detect this type of IRD is the focus of this study. In this study the expression **Polytomous Differential Item Functioning (PDIF)** was used to refer to this type of IRD analyses. The rationale underlying PDIF is essentially the same as DIF or DDF. PDIF consists of assessing if the groups under study are attracted differently to the alternatives of an item, when the ability of the individuals is controlled.

Finally, some researchers interpret IRD from a multidimensional perspective (Ackerman, 1991a, 1991b, 1992; Ackerman & Evans, 1993; Camilli, 1992; Junker & Stout, 1991; Oshima & Miller, 1991; Shealy & Stout, 1989, 1991). Theoretically, IRD can be defined as resulting from the distributional differences of an irrelevant trait (Lord, 1980 (p. 220); Oshima & Miller, 1991). Therefore, theoretically, IRD studies can be conceptualized as an issue of multidimensionality (Camilli, 1992). This type of

IRD has been referred to as **Multidimensional DIF (MDIF)** (Ackerman, 1991a, 1991b, 1992; Camilli, 1992). The use of a multidimensional model for the assessment of IRD is especially needed when the test under study is known to measure multiple abilities.

To better define IRD within the multidimensional framework, the true and nuisance ability dimensions need to be defined (Ackerman, 1991a, 1991b, 1992). The true ability is defined as some theoretical ability that the test intends to measure. From a multidimensional perspective there might be more than one true ability dimension. The nuisance ability is defined as the skills that the examinees are using to answer an item but which are not intended to be measured by the test. IRD under the multidimensional framework exists when the conditional distribution of the nuisance ability differs across the groups of interest (Ackerman, 1991a, 1991b, 1992). This rationale can also be extended to the analysis of PDIF. This is mentioned here because one of the methods proposed for the analysis of PDIF was developed within this framework.

Defining uniform and nonuniform PDIF

Mellenbergh (1982) has further characterized the nature of DIF as being **uniform** or **nonuniform**. An item is said to display uniform DIF when a subgroup shows constant

DIF against another group at all the ability levels. There are some situations where an item displays DIF at lower ability levels against one subgroup, while at higher ability levels it displays DIF against the other subgroup. In these cases DIF is said to be nonuniform. Therefore, nonuniform DIF can be simply defined as a significant ability-by-group interaction, while uniform DIF can be defined as a significant group main effect (Holland & Thayer, 1986; Mellenbergh, 1982).

As with DIF, the nature of PDIF can also be characterized as being uniform or nonuniform. In this study uniform and nonuniform PDIF are defined as follows:

Uniform PDIF is said to exist when the differences in the response distributions among groups are constant across the levels of ability. In other words, this depicts a group main effect.

Nonuniform PDIF exists when the differences in the response distributions among groups are not constant across the levels of ability. This depicts an ability by group interaction.

The definitions of uniform and nonuniform PDIF used in this study simply extend the DIF definitions to the PDIF situation. These definitions, however, differ from the Wilson, Spray, and Miller (1993) view of uniform and nonuniform PDIF. They view the number of score categories (or response alternatives) as another confounding variable that should be considered in the definition of uniform and

nonuniform PDIF. By doing this, they run the risk of misclassifying an item as having both uniform and nonuniform DIF, when in fact it only has uniform DIF. For example, if the amount of uniform DIF is not constant across the score categories, they would label this as being an indication of both uniform and nonuniform DIF. However, this is simply an indication that a group-by-score category interaction exists (i.e. uniform PDIF). This does not indicate that nonuniform PDIF exists, especially if nonuniform PDIF is defined in terms of its relationship with ability. If score category is included in the definitions of uniform and nonuniform PDIF, uniform PDIF would exist when either a group main effect or a group-by-score category interaction is significant. Nonuniform PDIF would exist when a group-by-ability interaction and a group-by-ability-by-score category is significant.

The need to study PDIF

It is acknowledged that the response pattern of the distractors is related to ability (Green, Crone, & Folk, 1989; Levine & Drasgow, 1983). Actually, including all the alternatives in the analysis of IRD can provide more and better information concerning cultural, social, or gender differences than can only the responses to the correct answer. As well as providing a more detailed analysis of

the items, the analysis of all the alternatives can also help increase our understanding of the functioning of the distractors by identifying the distractors which cause the IRD (Dorans, Schmitt, & Bleistein, 1992); Drasgow, Levine, & William, 1985; Thissen, Steinberg, & Fitzpatrick, 1989). This cannot, however, be accomplished if only the correct/incorrect options are analyzed because in dichotomizing the data into two options the pattern of responses to the distractors is lost. Similarly, this understanding of item performance cannot be accomplished if only the distractors are analyzed because the information contained in the correct alternatives is lost.

Support that some information is found in the distractors can be found in the work of Jacobs and Vandeventer (1970) and Siegel (1963). Siegel (1963) analyzed the pattern of errors on several items of Raven's Progressive Matrices and found that sex and age differences significantly explained the pattern of errors. However, no relationship was found between the pattern of errors and the total test score. Jacobs and Vandeventer (1970) in categorizing the errors of the items of Raven's Progressive Matrices into "superior" (the most correct option of the distractors) were able to show a relationship between the proportion of superior answers with the total test score. The contention that some information is found in the distractors of Raven's

Progressive Matrices test is further confirmed in the work of Thissen (1976) by means of test information analysis. Thissen (1976) compared the test information obtained from Birnbaum's dichotomous item response model with Bock's (1972) polytomous item response model for twenty difficult items of the Raven's Progressive Matrices test. The test information with the polytomous model yielded one third to nearly twice as much information for the lower half of the ability range, while for the upper half of the ability range there were no substantial differences in the test information curves. The overall measurement error variances were found to be 0.30 for the dichotomous model and 0.28 for the polytomous model. For these items the reliability was .72 for the polytomous model and .70 for the dichotomous model. The results of Thissen (1976) are similar to Bock (1972). Bock's comparison was done on twenty difficult items on a form of the Cooperative Reading Tests. His results indicated an increase in test information for the lower half of the ability range of one and a half to 2 times with the polytomous model. For the upper half of the ability range, he found that both models yielded approximately the same amount of information. The average measurement error variance of ability was 0.175 with the dichotomous model and 0.139 for the polytomous model. This polytomous model resulted in an increase in test reliability from .825 to .861. Similar results were

also reported by Sympson (1983) and Thissen and Steinberg (1984) on three difficult items and on four items from the Operational Military Accession test, respectively. Again, they found that for the lower half of the ability range the test information was increased by about 50%. However, Thissen and Steinberg (1984) found that the dichotomous model seemed to provide slightly more information for the upper half of the ability range. They explained that this difference may be an illusory effect and that the dichotomous model may simply overestimate the amount of information it contains. Therefore, these researchers seem to suggest that using all the information contained in an item provides a slight overall reduction in measurement error variance and it provides a considerable increase in the ability estimation for the lower half of the ability range. This further suggests that there is some information to be gained in including all the alternatives and that IRD would be better assessed if all the alternatives of an item were analyzed. Because more test information is gained for the lower half of the ability range, including all the alternatives in the IRD analysis would be more useful for these individuals (Green, Crone, & Folk, 1989, Thissen, 1976). Also, more information would be gained by including all the alternatives in the IRD analysis for difficult tests because the distractors are

chosen with larger frequencies and thus they would provide more information (Thissen, 1976).

Traditional item analysis, often carried out on multiple-choice items, provides detailed information that can help the item writers diagnose items that are not performing well (Wainer, 1989). With this knowledge the item writers can develop hypotheses about how to improve the item(s) (Thissen, Steinberg, & Fitzpatrick, 1989). Another benefit of multiple-choice item analysis is that an analysis of the distractors can provide some diagnostic clues in the evaluation of the functioning of items. Therefore, if item analysis studies can benefit from the analysis of all the alternatives, it is only normal to wish to extend these benefits to IRD studies.

Review of existing PDIF procedures

In this section, techniques which have been previously proposed for the analysis of polytomously scored items are presented. Some of the techniques presented here are simple extensions of the presently used DIF analyses to the case of polytomously scored items.

Log-Linear approach

Mellenbergh (1982) has suggested the general theory of log-linear models for assessing DIF. He showed that DIF can be assessed by analyzing a three-way contingency table

with log-linear models. Although, Mellenbergh (1982) applied the log-linear approach only to dichotomous response choices, he stated that the approach could also be extended to polytomous response choices. The application of log-linear models to polytomous response variables was done by Green, Crone, and Folk (1989). They applied the log-linear approach to the verbal portion of a recent Scholastic Aptitude Test. Their analysis included only the distractors as they found that the inclusion of the correct options tended to overwhelm the smaller differences observed among the distractors. There is no doubt that the inclusion of the correct alternative can somehow obscure the differences found in the distractors. However, this may not necessarily pose a problem because these small differences, although statistically significant, may not be practically significant. Also, the analysis of the distractors has the disadvantage of ignoring the differences which are found in the correct alternatives.

Veale and Foreman index

Veale and Foreman (1983) proposed the computation of a chi-square statistic on the distractors for each item as a method to assess if cultural variation is present. They define cultural variation in terms of the differences in the proportion of subjects in each group which are

attracted to each distractor. These differences in proportion are referred to as pull indices as they indicate the extent to which the distractors attract or pull the subjects belonging to different groups away from the correct option. The lack of cultural variation occurs when the pull indices are equal across the two groups. A major drawback of this method is that it does not control for the subject's ability. Veale and Foreman (1983) do acknowledge that cultural variation may be caused by a difference in group's ability; however they believe based on empirical data that this would not occur frequently. This assertion is questionable. Any IRD method should control for the differences in group abilities because there exists an interaction between the ability of the subjects and their choice of distractors (Levine & Drasgow, 1983).

Welch and Hoover procedures: The HW1, HW3, and the Mantel-Haenszel

Three procedures were compared by Welch and Hoover (1991) regarding their effectiveness in the analysis of polytomously-scored-items. The procedures that they compared were the Mantel-Haenszel procedure and two other statistics, namely the HW1 and HW3. The HW1 statistic is simply a series of independent t-tests computed at each

level of ability. The equation used at ability level k ($k=1,2,\dots,h$) is the following:

$$t_k = \frac{(\bar{X}_{1k} - \bar{X}_{2k}) - (\mu_{1k} - \mu_{2k})}{\sqrt{\frac{(S_{1k}^2 n_{1k} + S_{2k}^2 n_{2k})}{n_{1k} + n_{2k} - 2} \left(\frac{1}{n_{1k}} + \frac{1}{n_{2k}} \right)}} \quad (1)$$

with $df_k = n_{1k} + n_{2k} - 2$.

where

df_k = is the number of degree of freedom at ability k ;
 X_{1k} = is the sample mean of group 1 at ability k ;
 X_{2k} = is the sample mean of group 2 at ability k ;
 μ_{1k} = is the population mean of group 1 at ability k ;
 μ_{2k} = is the population mean of group 2 at ability k ;
 S_{1k}^2 = is the sample variance of group 1 at ability k ;
 S_{2k}^2 = is the sample variance of group 2 at ability k ;
 n_{1k} = sample size of group 1 at ability k ;
 n_{2k} = sample size of group 2 at ability k .

They use a significance test which combined the results of these series of t-tests. This test is given by

$$Z_c = \frac{\sum_{k=1}^h t_k}{\sqrt{\sum_{k=1}^h (df_k / (df_k - 2))}} \quad (2)$$

The statistic, Z_c , is then compared to the standard normal distribution in order to assess it's level of significance.

The HW3 is a weighted procedure of the HW1 statistic that more accurately represents the size of the two groups

under study. Using the HW1 statistic, the formula for the weighted procedure is

$$d_k = J_k \times \frac{n_{1k} + n_{2k}}{\sqrt{n_{1k} n_{2k}}} \times t_k, \quad (3)$$

where $J_k = 1 - (3(4m_k - 1))$, and $m_k = n_{1k} + n_{2k} - 2$.

If it is assumed that the d_k are independent and that SE_k are their standard errors. The weighted average of the k -independent effect sizes is

$$D_k = \frac{(d_1/SE_1^2) + \dots + (d_k/SE_k^2)}{(1/SE_1^2 + \dots + 1/SE_k^2)}. \quad (4)$$

The standard error of the weighted average, is given by

$$S_D = \frac{1}{\sqrt{1/SE_1^2 + \dots + 1/SE_k^2}}. \quad (5)$$

The method for testing the significance is then

$$Z = \frac{D_k}{S_D}. \quad (6)$$

This value of Z computed for the HW3 statistic is also compared to the standard normal distribution in order to assess it's level of significance.

The Mantel-Haenszel statistic that Welch and Hoover (1991) used is the generalization of this procedure to polytomous response variables. To understand the generalized formulation of the Mantel-Haenszel a general contingency table is described (Table 1).

Table 1: Illustration of a contingency table at ability level i where $i=1,2,\dots,k$

	Response Choice at ability i				Total
	1	2	...	g	
Y-score	Y_1	Y_2	...	Y_g	...
Group 1	A_{1i}	A_{2i}	...	A_{gi}	n_{1i}
Group 2	B_{1i}	B_{2i}	...	B_{gi}	n_{2i}
Total	M_{1i}	M_{2i}	...	M_{gi}	T_i

In Table 1, a Y-score is assigned to the various response choices. This is needed for the computation of the generalized Mantel-Haenszel procedure. Welch and Hoover (1991) have not specified how they computed this value, however some suggestions are found in Mantel (1963). For example, Mantel (1963) suggested that the rank of the response choice be used as the Y-score. In Table 1, at the i^{th} ability levels, there are A_{1i} subjects of group 1 who have selected the first response level and B_{1i} subjects of group 2 who have selected the first response. A summary

chi-square statistic, with one degree of freedom, is computed as follows:

$$\chi^2 = \frac{(\sum_i^k \sum_j^g A_{ji} Y_j - \sum_i^k E(\sum_j^g A_{ji} Y_j))^2}{\sum_i^k \text{var}(\sum_j^g A_{ji} Y_j)} , \quad (7)$$

where

$$\text{var}(\sum_j^g A_{ji} Y_j) = \frac{n_{1i} n_{2i}}{T_i^2 (T_i - 1)} [T_i \sum_j^g M_{ji} Y_j^2 - (\sum_j^g M_{ji} Y_j)^2] , \quad (8)$$

and

$$E(\sum_j^g A_{ji} Y_j) = \sum_i^k n_{1i} \frac{\sum_j^g M_{ji} Y_j}{T_i} . \quad (9)$$

To assess the effectiveness of these three procedures, ten polytomously scored items, having varying degrees of uniform differences, were generated using a partial credit model (Masters, 1984). Each of these ten IRD conditions were simulated under three equal sample size conditions (250, 500, 1000) and one unequal sample size condition (500 versus 1500). The first set of conditions were simulated with equal ability $N(0,1)$ distributions while the other conditions simulated unequal ability distributions with the distribution of the focal

group being non-normal ($N(-1,1)$). All the conditions were replicated 100 times and each item was analyzed for IRD using an external criterion (performance on a multiple choice test) as a conditioning measure.

Welch and Hoover (1991) Monte Carlo investigation revealed that the HW3 and the HW1 outperformed the Mantel-Haenszel across all the conditions. In addition, Welch and Hoover (1991) observed that the HW3 and the Mantel-Haenszel have consistent Type I error rates under the null hypothesis for all the conditions. The Mantel-Haenszel procedure was found to be less powerful when differences in ability existed. The Type I error rates seemed to be more affected for the conditions with unequal sample sizes and with unequal ability distributions.

There are, however, some implied assumptions associated with these techniques. The Generalized Mantel-Haenszel procedure used by Welch and Hoover (1991) requires that the response data be at least ordinal. The HW1 and the HW3 statistics, because of the calculation of group means for the t-statistics, appear to require that the data be measured on an interval scale. Therefore, all these techniques were not considered further because this study is concerned only with the analysis of nominal data.

Standardization approach

Dorans, Schmitt, and Bleistein (1992) presented a method of using the standardization approach with polytomous item responses. Their method simply consists of computing the standardized differences for each of the alternatives. Actually, most DIF procedures may be extended to the polytomous case by simply replacing the keyed response with the option of interest in the computation. If all the alternatives are of interest, this may be done for all the alternatives. This although adequate, does not provide an overall test of significance for all the items. Also, analyzing PDIF in this fashion would surely increase the possible number of Type I errors.

Logistic regression

Recently, Wilson, Spray, and Miller (1993) presented a method for identifying nonuniform DIF in polytomous items using the logistic regression method. They essentially analyzed nonuniform PDIF in a dichotomous manner. Instead of computing one logit function per item, they computed $J-1$ logit functions per item, where J equals the number of item categories. In their Monte Carlo investigation, they assessed the usefulness of three coding schemes in the detection of nonuniform PDIF, namely the continuation ratio logits, the cumulative logits, and

the adjacent categories. All these coding schemes required that the data be at least ordinal, because the responses were compared in an orderly fashion. The continuation ratio logits scheme artificially dichotomized the response categories in an increasing fashion (i.e. those that received a score of zero versus all others, one versus all others, and two versus all others). The cumulative coding scheme also compares the response categories in an ordered fashion - those who received a zero score were compared to all others, then the zero and one scores were compared to all others, and so on. Finally, the adjacent categories coding scheme compares score categories that are adjacent to one another. For example, the zero score category was compared with the one category, then the one score category was compared to the two category, and so on. Of these codings, the continuation ratio logits scheme is the only one which uses an orthogonal coding scheme. Therefore the J-1 regression results can be added to yield an overall omnibus test of significance, with J-1 degrees of freedom.

Wilson, Spray, and Miller (1993) employed Monte Carlo methods in order to investigate the effectiveness of the logistic regression procedure using the three different coding schemes in its' assessment of nonuniform PDIF. Data were simulated, using the generalized partial credit model (Muraki, 1991), with two different sample sizes (500 and

2000) and 25 simulated items. Four conditions of PDIF were studied and each of these conditions were replicated 100 times. The first three conditions of PDIF varied only the a or discrimination parameter in order to introduce nonuniform PDIF. The fourth condition varied only the b , or difficulty parameter. According to the authors, varying only the b parameter results in nonuniform DIF across the item as well as introducing uniform DIF within each of the score categories. Actually, what they simulated is uniform DIF within the score category that is not constant across the score categories. They, however, incorrectly labelled a difference across the score category as being an indication of a group-by-ability interaction (nonuniform DIF). Instead, this is an indication that there is a group-by-score category interaction. That distinction is important to understand because it affects the interpretation of their results.

Not surprisingly, their results showed that the larger sample size conditions were more powerful than the smaller one. Also, as stated by the authors, the continuation-ratio and adjacent-categories logit schemes were found to be less powerful because certain score categories were ignored on selected regressions (i.e. they were coded as missing for a specific regression). Wilson, Spray, and Miller (1993) found that all the coding schemes were able to adequately detect nonuniform PDIF introduced

by varying the a parameter. As for the b conditions, the authors state that all the approaches were not able to detect what they refer to as "the nonuniform DIF across the score categories". Instead, they found that all the approaches incorrectly detected this effect as uniform DIF. This is not surprising because in varying only the b parameter they have not introduced nonuniform DIF, but instead have introduced uniform DIF that is not constant across the score category. In other words, they have introduced a group-by-score category interaction and not a group-by-ability interaction. In failing to make that distinction the authors have used a test of a group by-ability-interaction (nonuniform DIF) to assess a group-by-score category interaction and incorrectly concluded that the approaches cannot detect nonuniform DIF that is introduced by varying the b parameter. Actually, their results indicate that all their approaches were able to adequately detect the uniform and nonuniform DIF simulated. The authors concluded otherwise simply because they failed to make the distinction between a group-by-ability interaction and a group-by-score category interaction.

IRT models

Thissen, Steinberg and Fitzpatrick (1989) and Thissen, Steinberg, and Wainer (1988) recognized that

multiple-choice item response models are potentially applicable in the detection and understanding of PDIF. There exists a variety of polytomous nominal models such as those developed by Bock (1972), Samejima (1979), Simpson (1983), Thissen (1976), and Thissen and Steinberg (1984) among others. The use of multiple-choice models is fairly difficult to apply routinely, primarily because the model estimation is complicated. The indeterminacies of multiple-choice models are eliminated by imposing some constraints on the model. Thissen, Steinberg, and Fitzpatrick (1989) found that with MULTILOG (a program for multiple-choice response model) there is a great deal of trial-and-error involved before selecting appropriate constraints on the model. Also, these models would probably require larger sample sizes than their dichotomous counterparts. Furthermore, multiple-choice models as implemented in MULTILOG require that the user specify an anchor test (i.e., a subset of items constrained to have equal parameters between groups), in order to test the hypotheses in PDIF (Thissen, Steinberg, & Wainer, 1993). As of now, the process of using multiple-choice IRT models is both labour intensive and computationally intensive, and therefore, these models are not ready to be implemented. Before we can hope to use polytomous nominal or ordinal IRT models for the analysis of PDIF more research is needed to assess the sample size

needed with these models to obtain adequate parameter recovery. Most importantly, a systematic approach should be developed to get the appropriate constraints on the model (Thissen, Steinberg, & Wainer, 1993).

Summary of the approaches presently used to assess PDIF

Of all the approaches presented in this section, the log-linear approach is the only one that appears to have some potential in the analysis of PDIF.

As mentioned earlier in this section, the HW1, HW3, Mantel-Haenszel, and logistic regression model as employed by Wilson, Spray, and Miller (1993) require that the response data be at least ordinal. Therefore, they are not appropriate for the analysis of multiple-choice data because they cannot be used with nominal data. It should be noted, that the logistic regression model as presented by Wilson, Spray, and Miller requires ordinal data; however, it is shown in the next chapter that this method can also be used for the analysis of PDIF with nominal response data.

The method proposed by Veale and Foreman (1983) is also inappropriate because it confounds impact with DDF.

The standardization procedure shown by Dorans, Schmitt, and Bleistein (1992) was not used in this investigation because the method does not provide a

general omnibus test in order to identify if the item is exhibiting PDIF.

Finally, item response models for nominal data have a potential use in the analysis of PDIF, however the techniques have not yet been adequately developed and studied.

Therefore, the log-linear analysis is the only one of the present techniques which was used in this study. Other approaches were employed in this study for the analysis of PDIF and are presented in the next chapter.

CHAPTER 3

Description of the PDIF Approaches Studied

In this chapter four new approaches, in addition to the Log-Linear (LOG) approach of Mellenbergh (1982), are presented for the analysis of PDIF. These approaches are the MANOVA (MCA), MANCOVA (MCO), Dual Scaling (DS), and the Polytomous Logistic Regression (PLR) approaches. An overview of the mathematical formulation of the MCO and PLR approaches is presented because these have, either not been presented before, or because the mathematical formulation is fairly unknown. As for the DS, MCA, and LOG, only the rationale of these procedures is presented. This is done because the mathematical formulation of the MCA approach, although in a different context than PDIF, has already been presented in Gessaroli (1985). The LOG approach has already been employed in the analysis of polytomous IRD by Green, Crone, & Folk, 1989. Finally, the mathematical formulation of dual scaling has been presented in Nishisato (1980).

Before presenting the approaches, a small numerical example is presented to allow the formulation of each technique from a common framework. Then, each approach is presented. Empirical evidence for the use of the approach is also given whenever it is available. Finally, this chapter concludes with a comparison of the approaches proposed for the investigation of PDIF.

Formulation of a general example

Let us assume that a subject in group k ($k=1,2,\dots,g$) responded to item i ($i=1,2,\dots,p$) by choosing response alternative m ($m=1,2,\dots,n_i$), where n_i is the total number of possible response alternatives for item i . The response-pattern matrix for N subjects is shown in Table 2. In Table 2, the chosen alternative is coded as 1 with the other options coded as 0. Each subject is asked to choose only one alternative for item i .

Table 2: A response-pattern matrix for a multiple-choice test

Subject	Group	Item 1				Item 2				...	Item p			
		1	2	...	n_1	1	2	...	n_2		1	2	...	n_p
1	1	1	0	...	0	0	1	...	0	...	0	0	...	0
.	.	.	.	...	.	.	.	...	.	...	.	.	...	.
.	1	1	0	...	0	1	0	...	0	...	0	1	...	0
.	2	1	0	...	0	1	0	...	0	...	0	1	...	0
.	.	.	.	...	.	.	.	...	.	...	.	.	...	.
.	2	0	0	...	1	0	1	...	0	...	1	0	...	0
.	.	.	.	...	.	.	.	...	.	...	.	.	...	.
.	.	.	.	...	.	.	.	...	.	...	.	.	...	.
.	g	0	1	...	0	0	1	...	0	...	0	0	...	0
.	.	.	.	...	.	.	.	...	.	...	.	.	...	.
N	g	1	0	...	0	0	1	...	0	...	0	0	...	0

Let us define the following notation:

- N = the total number of subjects;
- g = the number of groups;
- n_i = the number of options for item i , $i=1,2,\dots,p$;
- p = the total number of multiple-choice items.

The PDIF analyses used in this study were conditional on the ability of the subjects. This conditioning was needed because it is common to find that people with different abilities choose different alternatives (Holland & Thayer, 1986; Levine & Drasgow, 1983). Actually, omitting this conditioning results in statistical conclusions suggesting the deletion of items on the basis of differences in group performance that are not due to IRD (Green, 1991; Holland & Thayer, 1986; Kelderman & Macready, 1990; Linn & Drasgow, 1983; Mellenbergh, 1988). This effect is commonly referred to as impact (Holland & Thayer, 1986), where impact is a real group difference on the valid skill.

Conditioning can be defined by an independent measure of ability, if one is available, but often the total test score is used (Holland & Thayer, 1986). The latter approach was used in this study. Actually, two types of conditioning strategies were defined. The first approach consisted of using the total test score for conditioning, while the other strategy used the total test score to sort the subjects into ability groups. The choice between using the total test score or some ability group as the measure of ability depends on the nature of the PDIF approaches. Further details on the type of conditioning used in this study is provided when each of the PDIF approaches are presented.

The purpose of PDIF is to assess if persons having the same ability but belonging to different groups prefer different alternatives to a given item i . In the following section the techniques proposed in this study to assess PDIF are described using the above notation.

The MAN(C)OVA approaches

The origin of the MAN(C)OVA techniques presented here can be traced to the work of Hotelling (1933, 1936). Hotelling (1933) developed the method of canonical correlation to analyze the relationship between two sets of variates. More specifically, Hotelling (1933, 1936) developed a technique based on eigenvalue theory from which optimum linear combinations of the continuous variables can be derived. It is this approach that Gessaroli (1985) extended to the analysis of multiway contingency tables. The method presented by Gessaroli (1985) allows the analysis of multiway contingency tables through a Multivariate Analysis of Variance (MANOVA). In this study, the Gessaroli approach was employed for the analysis of PDIF. Gessaroli (1985) presented the case where two independent variables were measured on a categorical scale. However, all the independent variables do not need to be categorical. Accounting for ability with this method is achieved by either considering ability as a covariate (continuous) or as another independent variable

(categorical). Therefore, with this approach the ability can be measured using either the total test score or by some grouping based on the total test score. When the ability is treated as a covariate this approach is referred to as a MANCOVA or MCO approach, and when the ability is categorical the approach is referred to as a MANOVA or MCA approach.

The Gessaroli (1985) method of analyzing categorical dependent variables through a MANOVA procedure is relatively unknown. Therefore, an overview of the mathematical formulation procedure is presented in the next section. This is only done for the MCO approach since Gessaroli (1985) presented the mathematical formulation of this procedure for the MCA approach.

MANCOVA (MCO) approach. Employing the notation developed in the general example, let us also define the following for the MCO approach:

$$F = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ 1 & 0 & \dots & 0 \end{bmatrix} = \begin{array}{l} N \times n_i, \text{ the matrix of responses to} \\ \text{item } i. \end{array}$$

$$F^* = [y|F], \text{ the augmented matrix, where } y \text{ is an } N \times 1 \text{ vector of ability;}$$

f = the $n_i \times 1$ vector of column totals of F .

Using this notation, the unadjusted total sum of squares (SS_t) in a usual MANCOVA has been defined as follows:

$$\begin{aligned} SS_t &= x'F'F^*x - cf \\ &= x'[y|F]'[y|F]x - cf, \end{aligned} \quad (10)$$

where $cf = x'f'fx/N$, is the correction factor.

SS_t may be written in a matrix as follows:

$$SS_t = \begin{bmatrix} y'y & y'F \\ F'y & F'F \end{bmatrix}, \quad (11)$$

It should be mentioned that the correction factor is omitted from this matrix. In the above matrix, the upper left-hand partition represents the sum of squares for the covariate ($SS_t^{(y)}$). The lower right-hand matrix is the partition of the sum of squares and cross products for the dependent variables ($SS_t^{(F)}$). Finally, the two off-diagonal partitions contain the cross products of the covariate with the dependent variables ($SS_t^{(F'y)}$ and $SS_t^{(y'F)}$).

Each of these partitions which form the SS_t matrix are used to calculate the SS_t adjusted (SS_t^*) for the

effect of the covariate. In matrix notation, SS_t^* is given by

$$\begin{aligned} SS_t^* &= SS_t^{(F/F)} - SS_t^{(F/y)} (SS_T^{(y/y)})^{-1} SS_t^{(y/F)} - cf \\ &= X' [F'F - F'y(y'y)^{-1}y'F] X - cf. \end{aligned} \quad (12)$$

SS_{bet} , the unadjusted sum of squares between, in the traditional MANCOVA may be computed as follows:

$$\begin{aligned} SS_{bet} &= X'F'A_\alpha(A_\alpha'A_\alpha)^{-1}A_\alpha'F'X - cf \\ &= X'[y'F]A_\alpha(A_\alpha'A_\alpha)^{-1}A_\alpha'[y'F]X - cf, \end{aligned} \quad (13)$$

where $\alpha =$ is the effect to be maximized;
 A_α is the design matrix corresponding to the effect to be maximized.

The cross-products matrix of SS_{bet} without the correction factor is

$$SS_{bet} = \begin{bmatrix} y'A_\alpha(A_\alpha'A_\alpha)^{-1}A_\alpha'y & y'A_\alpha(A_\alpha'A_\alpha)^{-1}A_\alpha'F \\ F'A_\alpha(A_\alpha'A_\alpha)^{-1}A_\alpha'y & F'A_\alpha(A_\alpha'A_\alpha)^{-1}A_\alpha'F \end{bmatrix}, \quad (14)$$

In this matrix, the upper left-hand partition represents the between sum of squares for the covariate ($SS_{bet}^{(y)}$). The lower right-hand matrix is the partition of the between sum of squares and cross products for the dependent variables ($SS_{bet}^{(F)}$). The two off-diagonal partitions contain the between cross products of the

covariate with the dependent variables ($SS_{bet}^{(F'y)}$ and $SS_{bet}^{(y'F)}$). Again, each of these partitions are used to calculate the adjusted sum of squares between (SS_{bet}^*) as follows:

$$\begin{aligned} SS_{bet}^* &= SS_{bet}^{(F'F)} - SS_{bet}^{(F'y)} (SS_{bet}^{(y'y)})^{-1} SS_{bet}^{(y'F)} - cf \\ &= X' [F'AF - F'AY(y'AY)^{-1} y'AF] X - cf \end{aligned} \quad (15)$$

where $A = A_\alpha (A_\alpha' A_\alpha)^{-1} A_\alpha'$

Because the sums of squares in MANCOVA are related as follows:

$$SS_t^* = SS_{bet}^* + SS_w^* \quad (16)$$

SS_w^* is easily computed by taking the difference between SS_t^* and SS_{bet}^* , where SS_w^* is given by

$$\begin{aligned} SS_w^* &= SS_t^* - SS_{bet}^* \\ &= X' [F'F - F'y(y'y)^{-1} y'F - F'AF + F'AY(y'AY)^{-1} y'AF] X \end{aligned} \quad (17)$$

The goal of a MANCOVA is to calculate the optimum vector of weights (x) which maximizes the following ratio

$$\lambda = \frac{SS_{bet}^*}{SS_w^*} \quad (18)$$

λ is a ratio of two quadratic forms, that in it's present form cannot be solved because $F'F$ is singular. This occurs when all the alternatives (n_i) for item i are analyzed, it is possible to determine a subject's response by knowing only n_i-1 alternatives. Therefore, F needs to be reparametrized, into F_r , to eliminate the obvious relationship found among the dependent variables. This problem has been researched by several investigators (Healy & Goldstein, 1976, McDonald, Torii, & Nishisato, 1971; Nishisato, 1980). In this study, the reparametrization procedure suggested by Nishisato (1980) was used. The relation between F and F_r (F reparametrized) can be expressed in the general form

$$F_r = FJ' , \quad (19)$$

where J' is the matrix of contrasts among the dependent variables that consist of contrasting each of the n_i-1 alternatives with the last alternative of item i . An example of the J matrix is shown in Table 3, when item i has 4 alternatives.

Table 3: Example of the J matrix for item i

Alternatives of item i	contrast		
	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1
4	-1	-1	-1

This reparametrization also needs to be applied to x and f as follows:

$$x_r = xJ' , \quad (20)$$

and

$$f_r = fJ' \quad (21)$$

where x_r = is the vector of weights reparametrized.
 f_r = is the vector f reparametrized.

Substituting these reparametrizations into equations 15 and 17 yield the following equations:

$$\begin{aligned}
 SS_{bet}^r &= x_r' [F_r' A F_r - F_r' A y (y' A y)^{-1} y' A F_r] x_r - x_r' f_r f_r / N x_r \\
 &= x_r' [F_r' A F_r - F_r' A y (y' A y)^{-1} y' A F_r - f_r f_r / n] x_r \\
 &= x_r' B x_r
 \end{aligned} \quad (22)$$

where

$$B = F_r' A F_r - F_r' A y (y' A y)^{-1} y' A F_r - f_r f_r / n$$

$SS_{bet}^r =$ represents the adjusted between sums of squares for the reparametrized dependent variables;

and

$$\begin{aligned} SS_w^r &= x_r' [F_r' F_r - F_r' y (y' y)^{-1} y' F_r - F_r' A F_r + F_r' A y (y' A y)^{-1} y' A F_r] x_r \\ &= x_r' C x_r \end{aligned} \quad (23)$$

where

$$C = F_r' F_r - F_r' y (y' y)^{-1} y' F_r - F_r' A F_r + F_r' A y (y' A y)^{-1} y' A F_r$$

$SS_w^r =$ represents the adjusted within sums of squares for the reparametrized dependent variables.

Once these reparametrizations have been done, then a usual MANCOVA is conducted which consists of finding the vector of weights (x_r) that maximize the following ratio:

$$\lambda = \frac{SS_{bet}^r}{SS_w^r} = \frac{x_r' B x_r}{x_r' C x_r} \quad (24)$$

Essentially, the MCO approach is a typical MANCOVA that is conducted on the reparametrized dependent variables, and the vector of weights (x_r) that maximizes equation 24 are typically derived using traditional eigenvalue decomposition. Because the details of this procedure can be found in any basic multivariate textbook (e.g. Tatsuoka & Lohnes, 1988), it will not be outlined here.

In summary, this PDIF approach uses a traditional MANCOVA on the reparametrized dependent variables (i.e. response alternatives) in order to assess uniform and nonuniform PDIF. With this approach, nonuniform PDIF is assessed through a test of the homogeneity of the regression slopes, which essentially tests if there is a covariate (i.e. ability)-by-group interaction. If there is no such interaction, then one can proceed to test uniform PDIF by testing the equality of the adjusted group means. However, if a significant covariate-by-group interaction is found, one cannot follow the analysis by a test of the adjusted groups means as this test would be inappropriate (Stevens, 1986; Tabachnick & Fidell, 1989; Tatsuoaka & Lohnes, 1988).

MANOVA (MCA) approach. The MCA approach proposed for the analysis of PDIF is again a traditional MANOVA procedure that is conducted on the reparametrized dependent variables (i.e. response alternatives). The same reparametrization procedure as shown in Table 3 has to be performed on the dependent variables before the data can be subsequently analyzed with a traditional MANOVA (Gessaroli, 1985). The mathematical formulation of the MCA procedure is not presented in this section because it can be found in any multivariate textbook (for example in Bock, 1975; Tabachnick & Fidell, 1989; Tatsuoaka & Lohnes,

1988). In this approach the ability groups are considered to be different levels of an independent variable. Conditioning for ability with this technique is achieved by using a blocking approach (Bock, 1975; Tabachnick & Fidell, 1989). This essentially uses a sequential procedure in order to control for the lower order effects. To achieve proper conditioning, the order of the effects have to be entered in the following order; 1) ability main effect 2) group main effect and 3) ability-by-group interaction. Therefore, uniform PDIF, is tested by the group main effect when the ability differences have already been accounted for. Nonuniform PDIF, with this approach, is tested by assessing the group-by-ability interaction when all the other effects have already been accounted for.

Empirical support for the MCA and MCO approaches

The possibility of using canonical correlation for the analysis of categorical variables can be traced in the works of Fisher (1940) and Maung (1941). These studies have basically demonstrated how the canonical correlation, as defined by Hotelling, can be generalized to two-way contingency tables. This idea was further developed in the works of Kendall and Stuart (1967, pp.568-578), Lancaster (1958), Srikantan (1970) among others. Gigula (1986), Gigula and Haberman (1988), and Goodman (1979,

1981) have formulated a series of restricted canonical correlation models which allow for an in depth analysis of two-way contingency tables.

Another method based on canonical correlation has been proposed by Gessaroli (1985), and Zwick and Cramer (1986) for the analysis of categorical data. This method simply analyzes a multiway contingency table through MANOVA. In order to analyze contingency tables through a MANOVA approach, it is necessary to express the dependent variable in terms of I-1 indicator variables (Gessaroli, 1985; Zwick and Cramer, 1986). Zwick and Cramer (1986) described the use of this method for two-way contingency tables while Gessaroli (1985) showed the effectiveness of this method for three-way contingency tables.

Gessaroli (1985) studied the performance of Roy's Largest Root Criterion, Bartlett's approximation of Wilks Λ , and Fisher's modification of the F-ratio for use in the hypothesis of the MANOVA technique with categorical data. Twenty-seven experimental conditions were tested. The subjects were divided in two factors, namely A and B. Factor A had two, three, and five levels while factor B had two levels. The parameters were varied so that each subject responded to either two, five, or eight items having two, five, or eight options. In this study the correlation among the items was set to .50. The simulation included 500 replications having 120 subjects per

experiment. This study showed that only Roy's Largest-Root Criterion maintained good Type I error rates for all the conditions examined. It was suggested that under the conditions studied, Roy's Largest Root criterion was most effective with categorical data. The technique needs to be extended to the analysis of PDIF under the conditions in which a typical PDIF analysis would be performed. This was one of the goals of this study.

The use of the MCA or MCO approaches with categorical data may pose a problem as multivariate normality may be violated with such data. However, this does not seem to be a problem as O'Neil (1978a, 1978b) has noted that for an $n \times m$ contingency table the Pillai criterion does not seem to be affected. Actually, the distribution of the $N \Lambda_\alpha$ is the same as the asymptotic distribution as in the normal case, namely, the α^{th} eigenvalue of a central Wishart distribution (O'Neil, 1978a, 1978b, 1981).

Also, it has been demonstrated by Knapp (1978) that the chi-square test of independence, among other statistical techniques, can be treated as a special case of canonical correlation analysis. This should not be too surprising because Kshirsagar (1972) demonstrated that the Pillai-Bartlett trace (λ) is a simple function of the Pearson χ^2 statistic when used in the analysis of two-way contingency table. The Pearson χ^2 can be computed from the Pillai-Bartlett as $\chi^2 = N\Lambda$ (Kshirsagar, 1972). In addition

to this equivalence, Holland and Levi (1980) noted that canonical correlation yields more information about the relationship in a two-way contingency tables than does the chi-square analysis.

The Dual Scaling (DS) approach

This approach uses Dual Scaling (or Optimal Scaling) to assess PDIF. More specifically, this approach suggests that the item responses may be scaled (i.e. quantified), using Dual Scaling, so that they may subsequently be analyzed with an ANOVA. This section explains how Dual Scaling was used to scale the item responses. Dual Scaling is a family of techniques which are mainly intended for the analysis of categorical data. The basic idea of Dual Scaling, according to Nishisato (1980), can be traced to the work of Guttman (1941, 1966), Hirschfeld (1935), Horst (1935), and Richardson and Kuder (1933). This study focused on the application of Dual Scaling to multiple-choice data which is discussed by Nishisato (1980).

Dual scaling presents a way of quantifying the nominal item responses so that the data may subsequently be analyzed with an ANOVA to assess PDIF. In order to understand how Dual Scaling might serve this purpose, the analysis of PDIF has to be viewed from a multidimensional perspective. To simplify the discussion, PDIF will be examined from a two-dimensional perspective. In the PDIF

analysis, it is assumed that the first dimension of a supposedly unidimensional test measures the intended ability and the second dimension measures what Ackerman (1991a, 1991b, 1992) terms the nuisance ability. The nuisance ability dimension is defined as a trait that examinees are using to solve a particular item but was not intended to be assessed (Ackerman, 1991a, 1991b, 1992). In this case PDIF occurs when the groups of interest differ in their underlying distribution on the nuisance ability.

Dual Scaling provides a method that can give estimates of these two dimensions, at least from an eigenvalue decomposition perspective. It computes the first dimension by assigning weights to the item options and subjects so as to maximize the correlation between the set of items and the subjects who responded to these items (Nishisato, 1980). These weights are calculated to produce options weights and subject weights that are normally distributed (Kendall & Stuart, 1967, p.589).

Generally, with a well designed test, this dominant dimension should represent the subject's ability. The second dimension is computed from the remaining variance, therefore the new weighting system is independent of the first dimension. It is actually only the item response weights of the second dimension that are of interest, because PDIF should occur when the groups of interest differ in their item weights on the nuisance ability. The

option weights obtained from the second dimension represent the amount of the nuisance dimension, after partialling out the effect of the subject's ability (the first dimension). Thus, an ANOVA using the second dimension weights corresponding to the chosen alternative as the dependent variables should provide a test of equality among groups on the nuisance dimension after controlling for ability. An ANOVA, testing for group differences on the second dimension's weight, yields only a test of uniform PDIF because only a test of equality among groups is possible. This approach, however, offers the possibility of analyzing uniform PDIF from a multidimensional perspective.

The Polytomous Logistic Regression (PLR) approach

The logistic regression procedure is a technique that is similar to multiple regression, but it is not bound by the assumptions of normality. The dependent variable is qualitative and the independent variables can be either qualitative or quantitative (Aldrich & Nelson, 1984; Fienberg, 1980; Hosmer & Lemeshow, 1989; Swafford, 1980). This technique will only be described when the total test score serves as the measure of ability. Using grouped ability levels reduces this technique to a log-linear model. The log-linear approach is discussed in the next section.

Logistic regression models have already been suggested to study differences between groups with dichotomous item responses (Bennet, Rock, & Kaplan, 1987; Swaminathan & Rogers, 1990). However, this study was concerned with the formulation of the logistic regression model for polytomous item responses. The mathematics of polytomous logistic regression models does exist (Aldrich & Nelson, 1984; Bock, 1975; Fienberg, 1980; Hosmer & Lemeshow, 1989; Swafford, 1980). However, polytomous logistic regression has not yet been used to study group differences with polytomous item responses. Actually, applications of polytomous logistic regression models are relatively new. The polytomous logistic regression model is a generalization of the dichotomous dependent variable logit model (Aldrich & Nelson, 1984; Fienberg, 1980; Hosmer & Lemeshow, 1989; Swafford, 1980). The presentation of the polytomous model is facilitated by first introducing the dichotomous model.

The dichotomous logistic regression model for DIF using the notation of the general example is

$$\ln \left[\frac{P(U_{jk}=1|\theta_{jk})}{P(U_{jk}=0|\theta_{jk})} \right] = \beta_{0k} + \beta_{1k}\theta_{jk} , \quad (25)$$

where

U_{jk} = the response of subject j ($j=1,2,\dots,N$) in group k ($k=1,2,\dots,g$) to an item;

β_{0k} = the intercept parameter for group k ;
 β_{1k} = the slope parameter for group k ;
 θ_{jk} = the ability of subject j in group k .

Solving for the probability that U_{jk} equals 1, results in equation 26.

$$P(U_{jk}=1|\theta_{jk}) = \frac{\exp(\beta_{0k} + \beta_{1k}\theta_{jk})}{1 + \exp(\beta_{0k} + \beta_{1k}\theta_{jk})} \quad (26)$$

From equation 25 it is shown that the dichotomous logistic regression model is parametrized in terms of the logit of the correct option ($U_{jk} = 1$) versus the incorrect options ($U_{jk}=0$).

In the polytomous case an approach similar to equation 25 is used. The polytomous responses are parametrized into pairwise n_i-1 logit functions (Aldrich & Nelson, 1984; Fienberg, 1980; Hosmer & Lemeshow, 1989; Swafford, 1980). In theory, any pairwise comparisons can be used. If item i has three alternatives the two logit functions could be

$$f_{(1)} = \ln\left[\frac{P(U_{jk}=1|\theta_{jk})}{P(U_{jk}=3|\theta_{jk})}\right] = \beta_{10k} + \beta_{11k}\theta_{jk} \quad (27)$$

and

$$f_{(2)} = \ln\left[\frac{P(U_{jk}=2|\theta_{jk})}{P(U_{jk}=3|\theta_{jk})}\right] = \beta_{20k} + \beta_{21k}\theta_{jk} \quad (28)$$

where

$f_{(x)}$ = the logit function (x) , $(x)=(1, \dots, n_i-1)$.
 β_{10k} = the first logit function intercept parameter for group k;
 β_{11k} = the first logit function slope parameter for group k;
 β_{20k} = the second logit function intercept parameter for group k;
 β_{21k} = the second logit function slope parameter for group k.

The general expression for the logit function with three alternatives is

$$f_{(x)} = \ln \left[\frac{P(U_{jk}=e|\theta_{jk})}{P(U_{jk}=E|\theta_{jk})} \right] = \beta_{e0k} + \beta_{e1k} \theta_{jk} , \quad (29)$$

where E is equal to one of the alternatives $(1, 2, \dots, n_i)$;
 e is different than E and $e=(1, 2, \dots, n_i)$.

As in the dichotomous case, solving for the probability that U_{jk} equals outcome e is

$$P(U_{jk}=e|\theta_{jk}) = (\exp^{(\beta_{e0k} + \beta_{e1k} \theta_{jk})}) * (P(U_{jk}=E|\theta_{jk})) . \quad (30)$$

The $P(U_{jk}=E|\theta_{jk})$ is known because

$$\sum_{m=1}^{n_i} P(U_{jk}=m|\theta_{jk}) = 1 . \quad (31)$$

Therefore,

$$\begin{aligned}
 P(U_{jk}=E|\theta_{jk}) &= \frac{1}{1 + \sum_{x=1}^{n_j \neq E} \exp(f_x)} \\
 &= \frac{1}{1 + \sum_{\theta=1}^{n_j \neq E} \exp(\beta_{e0k} + \beta_{e1k}\theta_{jk})} .
 \end{aligned} \tag{32}$$

Substit

uting into equation 30, the probability the U_{jk} equals alternative e is

$$P(U_{jk}=e|\theta_{jk}) = \frac{\exp(\beta_{e0k} + \beta_{e1k}\theta_{jk})}{1 + \sum_{\theta=1}^{n_j \neq E} \exp(\beta_{e0k} + \beta_{e1k}\theta_{jk})} . \tag{33}$$

In comparing expressions 33 and 26, it is observed that the difference between the polytomous model and the dichotomous model is found in the denominator. By examining equation 33, it becomes obvious that the polytomous model reduces to the dichotomous model when there is only one logit function. Therefore, one can see that the polytomous model is a straightforward extension of the dichotomous logit model and that no new assumptions are added (Aldrich & Nelson, 1984).

The denominator of expression 33 is simply a normalizing factor which ensures that the probability terms add up to one. In expression 33, the normalization depends on the selection of a baseline (E). In equation 33

E corresponds to any alternative of item i . In this study E was always the last alternative of the item.

With this approach, the alternatives of an item show PDIF if subjects having the same ability but belonging to different groups have different probabilities of success on the alternatives. It follows that no PDIF is present if the intercepts and slopes are the same for the two groups for each logit function. If one of the intercepts differ between the groups and the slopes are the same on each logit, then uniform PDIF is present for that item. However, if the intercepts differ between the groups nonuniform PDIF is present for that item.

The parameters of the polytomous model and their asymptotic standard errors were estimated from the sample using the maximum likelihood method. The likelihood ratio χ^2 served to evaluate the model. Uniform and nonuniform PDIF were assessed by taking the difference in the likelihood ratio χ^2 .

Empirical support for the polytomous logistic regression approach

Polytomous logistic regression as proposed in this study has not yet been used for the analysis of PDIF. It has, however, been employed in the analysis of DIF to study group differences (Bennet, Rock, & Kaplan, 1987) and compared with the Mantel-Haenzsel procedure (Swaminathan &

Rogers, 1990). In their study, Swaminathan and Rogers (1990) simulated six conditions having two levels of sample size (250 and 500 per group) and three levels of test length (40, 60, and 80). For each test, they simulated 20% of the items as having DIF with half of this DIF being uniform and the other half being nonuniform. The nonuniform DIF simulated in Swaminathan and Rogers (1990) study was disordinal interaction between ability and group membership, meaning that the characteristic curves cross in the middle of the ability range. Their results indicated that the logistic regression technique was more powerful than the Mantel-Haenzsel technique in detecting nonuniform DIF and as powerful in detecting uniform DIF. It is possible that with nonuniform DIF that is not disordinal the differences among the techniques may vary (Swaminathan & Rogers, 1990).

The study of Wilson, Spray, and Miller (1993) which has been presented in detail in Chapter 2, also lend some support for the use of the PLR procedure. Although, they did not used a polytomous method for the analysis of PDIF, their methods were still successful in the detection of PDIF that was introduced by varying the a or b parameters.

The log-linear (LOG) approach

Log-linear analysis is the usual extension of the chi-square test for two-way contingency tables to higher order tables. Basically, the technique consists of transforming the classic multiplicative model to a linear model (Baker, 1981; Baker & Subkoviak, 1981; Fienberg, 1980; Fienberg & Meyer, 1983; Haberman, 1978; Swafford, 1980). The log-linear analysis requires that all the variables be categorical. Therefore, ability groups constructed from the total test score serve as the ability measure. An illustration of the log-linear technique using a three-way contingency will be briefly discussed. As the technique has already been used by Green, Crone, and Folk (1989) the formulation is not discussed, but instead the rationale of the technique will be emphasized. The three-way contingency table has 1 ability levels with g groups and n_i alternatives for item i. The log-linear approach in this study differs from that in Green, Crone, and Folk (1989) only in that all the alternatives of the item are studied instead of looking only at distractors.

Let us first define the saturated log-linear model for a three-way contingency table. The saturated model is

$$\ln F_{(jkm_1)} = U + U_{A(j)} + U_{G(k)} + U_{Rm_1} + U_{AG(jk)} + U_{AR(jm_1)} + U_{GR(km_1)} + U_{AGR(jkm_1)} \quad (34)$$

where

F = is the frequency of the jkm 's cell, where
 $j=1,2,\dots,l$, $k=1,2,\dots,g$, and $m=1,2,\dots,n_i$;
 U = is a constant frequency in all cells;
 U_A = is the contribution to the cell frequency for the
ability factors, where l is the total number of
ability groups;
 U_G = is the contribution to the cell frequency for the
group factor, where g is the total number of groups;
 U_R = is the contribution to the cell frequency for the
alternative factor of item i , where n_i is the total
number of alternatives to item i ;
 U_{AG} = is the contribution to the cell frequency for the
ability by group interaction;
 U_{AR} = is the contribution to the cell frequency for the
ability by alternative interaction;
 U_{GR} = is the contribution to the cell frequency for the
group by alternative interaction;
 U_{AGR} = is the contribution to the cell frequency for the
ability by group by alternative interaction.

Employing a simplified notation the saturated model can be written as follows:

Model 1: $[A], [G], [R], [AG], [AR], [GR], [AGR]$.

For the PDIF analysis, like the DDF analysis of Green, Crone, and Folk (1989), and the DIF analysis of Mellenbergh (1982) two other models must be computed.

These are

Model 2: $[A], [G], [R], [AG], [AR], [GR]$,

and

Model 3: $[A], [G], [R], [AG], [AR]$.

The parameters of the log-linear models as well as their asymptotic standard errors are estimated through

Maximum Likelihood (Baker, 1981; Baker & Subkoviak, 1981; Fienberg, 1980; Fienberg & Meyer, 1983; Haberman, 1978; Mellenbergh, 1982; Swafford, 1980). The fit of the model can be assessed using the Pearson Chi-Square (χ^2) or the G^2 -statistic. As the log-linear model is based on the Maximum Likelihood estimates, the G^2 -statistic is more appropriate than the Pearson chi-square (Baker, 1981; Baker & Subkoviak, 1981). One of the interesting properties of the G^2 -statistic is that for two hierarchically nested models the difference between the G^2 -statistics is asymptotically chi-square distributed with the degrees of freedom equal to the difference in degrees of freedom between models (Fienberg, 1980; Haberman, 1978; Mellenbergh, 1982).

Mellenbergh (1982) stated that nonuniform DIF can be assessed by taking the difference between model 1 and model 2. This results in a test of the interaction of ability by group by alternatives, thus depicting nonuniform PDIF. The difference between model 2 and model 3 results in a test of the interaction of group by alternatives. The test of this interaction corresponds to a test of uniform PDIF.

Relationship among the PDIF approaches studied

Greenacre (1984), and Tenenhaus and Young (1985) demonstrated that canonical correlation and dual scaling

share the same numerical procedure for their respective solutions. They, however, originated under different rationales and are given different interpretations. Canonical correlation and dual scaling were proven to be formally identical in two-way contingency tables (de Leeuw, 1971 in van der Heijden, de Leeuw, 1985; Nishisato, 1980).

In two-way contingency tables, Goodman (1979, 1981) showed under which conditions canonical correlation and dual scaling results are similar to those of the log-linear model. Goodman (1981) showed that the R-C (Row and Column) association model is actually a log-linear interaction model of the multiplicative form. This R-C association model, when the two-way frequency tables are generated from an underlying discrete bivariate normal distribution, has estimates that are approximately equal to those from canonical correlation and dual scaling (Goodman, 1979, 1981; van der Heijden, de Falguerolles, & de Leeuw, 1989; van der Heijden & de Leeuw, 1985). Goodman (1979, 1981) formulated all the models using maximum likelihood.

Having provided the relationship between log-linear, dual scaling, and canonical correlation for two-way contingency tables, it is worthwhile to examine how logistic regression relates to the other techniques. There exists a close relationship between log-linear and

logistic regression models (Baker, 1981; Baker & Subkoviak, 1981; Fienberg & Meyer, 1983; Mellenbergh, 1982; Swafford, 1980). Distinguishing between explanatory and response variables helps the comparison between the log-linear and the logistic regression models (Baker, 1981; Baker & Subkoviak, 1981). Log-linear models define the contingency table in terms of response variables (i.e. dependent variables) only. Logistic regression models, on the other hand, define a contingency table in terms of response variables and explanatory variables (i.e. independent variables). Actually, the log-linear model can be thought of as being a special case of logistic regression in which all the response variables are categorical. Therefore, when all the response variables are categorical it is possible to show that a logistic regression model can be represented by an equivalent log-linear model (Fienberg & Meyer, 1983; Swafford, 1980). From the relationship between log-linear analysis and logistic regression one can easily see how logistic regression, canonical correlation and dual scaling are related in the analysis of two-way contingency tables.

The relationship among the four approaches studied is not as obvious in this study because higher order contingency tables were analyzed. Also, the nature of the ability scale was not constant among the approaches. The LOG and MCA approach treated the ability variable as being

discrete, and the PLR and the MVO defined ability as being continuous. In contrast, the DS approach does not treat ability as being either continuous or discrete. Instead, ability is considered to be a dimension of the test. Conditioning on ability is obtained by eliminating the effect of the first dimension. More importantly, the dual scaling methodology was used only to scale the item responses, but PDIF was assessed with an ANOVA. Similarly, DS approach was computed based on a different contingency table than the MCA and LOG approaches. The grouping effect was not considered in the computation of DS. The DS method as used in this study, therefore, is not comparable to the Dual Scaling approach used by Goodman (1979, 1981) van der Heijden, de Falguerolles, and de Leeuw (1989), and van der Heijden and de Leeuw (1985). Thus, the relationships earlier mentioned between the Dual Scaling methodology with those of the MANOVA applied to categorical dependent variables and log-linear methodology to the analysis of two-way contingency tables are not applicable in this study. Actually, the DS methodology used in this study is comparable to the one used by Nishisato (1971).

Therefore, it is not known how the DS, PLR, and the MCO approaches will compare to other approaches. It is plausible to think that the different approaches that treat ability as continuous would be more powerful in the detection of PDIF (Swaminathan & Rogers, 1990).

CHAPTER 4

Methods

The procedures used to assess the performance of each approach are outlined in this chapter. Because the rationale underlying these approaches was already presented in Chapter 3, the approaches studied are only briefly reviewed. Then, the type of conditioning used for each approach is presented. This is followed by a presentation of the conditions under which these statistical approaches were compared. The procedures followed to generate the item responses are then outlined. The program used to generate the item responses is discussed, as well as the software used to compute each statistical procedure. Then, the uniform and nonuniform tests of hypotheses for the different approaches investigated are contrasted. Next, a presentation of the statistical procedures that were used to analyze the Monte Carlo results is presented. Finally, this chapter ends with a discussion of the limitations of this study.

The PDIF approaches

Five PDIF approaches were investigated in this study - a MANOVA and a MANCOVA approach applied to categorical dependent variables, an ANOVA approach based on the alternatives quantified with Dual Scaling, a Polytomous

Logistic Regression analysis and, finally, a Log-Linear analysis.

Conditioning strategies

All of the methods in this study were conditional on a measure of ability. Conditioning was needed because it is important to dissociate impact from PDIF (Holland & Thayer, 1986) in any IRD studies.

Two measures of observed ability were defined in this study, one being the total test score (continuous) and the other being ability groups (categorical) constructed from the distribution of the total test score. In this study, the statistics that were computed with ability groups were based on four levels of ability. Each level of ability contained approximately 25% of the subjects.

The type of conditioning used depended on the nature of the PDIF approach employed. The MAN(C)OVA was computed using both measures of ability. When the measure of ability was continuous the approach was a MANCOVA, which in this study is referred as continuous MANOVA (MCO). When the ability measure was categorical, the approach used was a MANOVA. The term employed in this study was categorical MANOVA (MCA). The Polytomous Logistic Regression (PLR) model was used only with a continuous measure of ability, because, as mentioned earlier, with ability groups this method reduces to a log-linear model. The Log-Linear (LOG)

model is, of course, limited to using only a categorical measure of ability.

With the Dual Scaling approach (DS), conditioning was automatically achieved by analyzing the second dimension weights. Therefore, no additional conditioning on ability was needed with this technique.

Thus, there were in total five statistics that were calculated from one set of data; two for the MAN(C)OVA approaches (MCA and MCO), one for the Polytomous Logistic Regression analysis (PLR), one for the Log-Linear analysis (LOG), and one based on the analysis of Dual Scaling (DS).

Simulation conditions

In this section, the conditions under which the PDIF statistics were computed are described. The factors that were manipulated in this study were test length, sample size, item difficulty, and the amount and location of PDIF. Other factors such as the number of groups, the ability of the groups, the number of items exhibiting PDIF, and the number of replications were held constant for all the conditions studied. The factors that were held constant in this study are described first.

Factors held constant

Number of groups. The statistics were compared for conditions where the number of groups was equal to two

($k=2$). In DIF the group of primary interest is referred to as the **focal group** and the group which is used as the standard to compare the performance of the focal group is referred to as the **reference group** (Holland & Thayer, 1986). This terminology was also employed in this study.

Ability distribution. It is important to assess the effectiveness of the PDIF statistics when the populations differ in their ability distributions. However, due to some practical constraints on the number of conditions, this was not considered in this investigation. In all the conditions simulated, the underlying ability of the focal and reference groups was equal and normally distributed ($N(0,1)$).

Number of items exhibiting PDIF. Mellenbergh (1982) suggested that each item be analyzed separately for DIF. The approaches studied in this investigation all can analyze PDIF for individual items. In this study the performance the PDIF methods was investigated with respect to only one item of the test.

Replications. In each condition, 500 samples of examinees were randomly selected for the focal and reference groups. On each sample, PDIF analyses were carried out using the PDIF approaches described earlier.

Factors manipulated

Test length. Test length affects the accuracy of the total test score as a measure of ability. Longer tests generally yield more reliable test scores (Swaminathan & Rogers, 1990). Because four of the statistics use the total test score or a grouping based on the total test score for conditioning, it is believed that test length might influenced the effectiveness of these statistics. Therefore, it is important to assess how the PDIF approaches performed with various test lengths. Response strings for test lengths of 25 and 50 items were generated. In all cases there were four alternatives per item (one correct alternative and three distractors).

Sample size. The size of the samples and the ratio between the samples both affect the robustness of the statistics through its effect on estimation (Swaminathan & Rogers, 1990). In a real testing situation the reference group is usually much larger than the focal group. Therefore, the performance of the statistics was assessed for unequal sample sizes. In reality the ratio between the focal and reference groups may vary greatly, however in this study, the ratio between these two groups was set at 2:1. The specific sample sizes which were used in this study were (1) 1500/750 and (2) 500/250.

Item difficulty. The difficulty of the item may affect the accuracy of the techniques to assess PDIF by varying the response distributions. In this study two item difficulties were investigated: (1) Difficulty of approximately 0.40, and (2) difficulty of approximately 0.70.

Amount and location of PDIF. Two amounts of PDIF were used: (1) NO PDIF and (2) PDIF. It was important to consider the NO PDIF condition because such an investigation can ascertain the nominal Type I error rates for the statistics under study. The NO PDIF groups were generated with the same input parameters (difficulty and discrimination parameters) but used a different starting seed for each group. This process is described later in this chapter. The initial input parameters (difficulty and discrimination parameters) for the NO PDIF conditions are found in Tables 4 and 5 for the 0.70 and 0.40 conditions, respectively. In Tables 4 and 5, the difficulty parameter is expressed in term of proportions (i.e. the proportion of subjects who choose alternative e), while the discrimination parameter is actually the biserial correlation of the alternative with the total test score.

Table 4: Proportion (difficulty parameter) and biserial correlation (discrimination parameter) for the 0.70 PDIF conditions

Amount of PDIF	Options	Difficulty parameter		Discrimination parameter
		Focal group	Ref. group	Focal and Ref. groups
NO PDIF	1	.1269	.1269	-.2596
	2	.1115	.1115	-.2447
	3	.0464	.0464	-.3467
	4*	.7152	.7152	.3953
PDIF 1	1	.1269	.1269	-.2596
	2	.1615	.1115	-.2447
	3	.0464	.0464	-.3467
	4*	.6652	.7152	.3953
PDIF 2	1	.1436	.1269	-.2596
	2	.1282	.1115	-.2447
	3	.0630	.0464	-.3467
	4*	.6652	.7152	.3953

* This indicates the correct alternative.

Table 5: Proportion (difficulty parameter) and biserial correlation (discrimination parameter) for the 0.40 PDIF conditions

Amount of PDIF	Options	Difficulty parameter		Discrimination parameter
		Focal group	Ref. group	Focal and Ref. groups
NO PDIF	1	.2786	.2786	-.2711
	2	.1207	.1207	-.5250
	3	.1889	.1889	-.1486
	4*	.4118	.4118	.6082
PDIF 1	1	.3286	.2786	-.2711
	2	.1207	.1207	-.5250
	3	.1889	.1889	-.1486
	4*	.3618	.4118	.6082
PDIF 2	1	.2953	.2786	-.2711
	2	.1374	.1207	-.5250
	3	.2055	.1889	-.1486
	4*	.3618	.4118	.6082

* This indicates the correct alternative.

PDIF was introduced by making the proportions in the alternatives of the item for the focal group different from those of the reference group. That difference was set so that the maximum difference in the alternatives corresponded to a difference in proportion of .05. A very small difference was chosen in order to be able to compare the power of the techniques in detecting PDIF.

Changing the proportion of the alternatives actually translates into changing the difficulty parameter of the alternatives. These item difficulties were then included into the mathematical model used to generate the item responses. The procedures followed to obtain the input parameters (difficulty and discrimination parameters) for the mathematical model used to generate the data is explained in detail later in the chapter.

Differences between the alternatives were generated in two ways, namely PDIF 1 and PDIF 2. These conditions are also depicted in Tables 4 and 5, for the 0.70 and 0.40 item difficulty conditions, respectively. As shown in Tables 4 and 5, the PDIF 1 condition consisted of making the probability of choosing the correct alternative harder for the focal group by .05. This was offset by setting one of the distractors to be more attractive for that group by .05. In the PDIF 2 condition the correct alternative was also made harder for the focal group by decreasing the probability by .05. However, this increase in the

difficulty of the correct option was compensated by increasing the probability of choosing each of the other options by $.05/3$.

Final design

In the NO PDIF case, data sets in each of this $2 \times 2 \times 2$ design (test length by sample size by item difficulty) were replicated 500 times. This yielded a total of 4000 NO PDIF data sets that were analyzed. Similarly for the PDIF case, data sets in each of this $2 \times 2 \times 2 \times 2$ design (test length by sample size by item difficulty by location of PDIF) were replicated 500 times. This yielded 8000 PDIF data sets. All the data sets (12000) were, in turn, analyzed with the five statistical approaches.

Generation of the item responses

In this section, the two-parameter polytomous logistic regression model employed to generate the data is described. The strategy followed to obtain the input parameters (difficulty and discrimination parameters) for the generation of the item responses is then explained. Finally, the procedures used to generate the item responses for the focal and reference groups are described.

The two-parameter polytomous logistic regression model

The item responses for each group were simulated using a two-parameter polytomous logistic regression model (Aldrich & Nelson, 1984; Bock, 1975; Hosmer & Lemeshow, 1989). The two-parameter polytomous logistic regression model used to assess the probability of getting alternative e to item i has already been described by equation 33 in Chapter 3. Note that the only difference between equation 33 and 35 is that the subscript i was added. This subscript was added to indicate that this equation was used for each item. This equation may be rewritten in the following way:

$$P_{e_i}(U_{jk}=e_i|\theta_{jk}) = \frac{\exp^{Z_{e_i k}}}{1 + \sum_{e_i=1}^{n_i-1} \exp^{Z_{e_i k}}} , \quad (35)$$

where

$$Z_{e_i k} = \beta_{e_i 0 k} + \beta_{e_i 1 k} \theta_{jk} , \quad (36)$$

P_{e_i}	=	The probability of choosing alternative e ($e=1,2,\dots,n_i-1$) to item i ($i=1,2,\dots,p$);
U_{jk}	=	the response of subject j ($j=1,2,\dots,N$) in group k ($k=1,2$);
θ_{jk}	=	ability of subject j in group k ;
$\beta_{e_i 0 k}$	=	the intercept parameter of alternative e to item i for group k ;
$\beta_{e_i 1 k}$	=	the slope parameter of alternative e to item i for group k ;
E	=	the last alternative of item i ($E=n_i$);
e	=	the alternatives of item i except the last one, ($e=1,2,\dots,n_i-1$).

Because

$$\sum_{m=1}^{n_i} P_m = 1 , \quad (37)$$

the probability that subject j in group k will choose alternative E to item i is given by

$$P_{E_i}(U_{jk}=E_i|\theta_{jk}) = 1 - \sum_{e=1}^{n_i-1} P_{e_i} \quad (38)$$

The parameters of the logistic regression model can alternatively be rewritten in terms of the a or discrimination parameter and b or difficulty parameter which are used in Item Response Theory (IRT) to characterize the item. This relationship has been shown by McDonald (1985, p.207) in the context of factor analysis. Rewriting expression (36) in terms of the IRT parameters leads to the following equation

$$Z_{iek} = a_{iek}(\theta_{jk} - b_{iek}) . \quad (39)$$

Then a substitution of equation (39) into equation (35) results in the following:

$$P_{e_i}(U_{jk}=e_i|\theta_{jk}) = \frac{\exp^{(a_{e_ik}(\theta_{jk}-b_{e_ik}))}}{1 + \sum_{e_1=1}^{n_i-1} \exp^{(a_{e_1k}(\theta_{jk}-b_{e_1k}))}} , \quad (40)$$

where

a_{eik} = the discriminating power of alternative e to item i for group k;
 b_{eik} = the difficulty parameter of alternative e to item i for group k.

The item responses were simulated using equation (40). Before outlining this procedure, the method followed to obtain the input parameters a_{eik} and b_{eik} is described in the next section.

Item parameters used to generate the data

The item parameters a_{eik} and b_{eik} used to generate the item responses were based on those found in the analysis of an actual test. In order to derive these item parameters, the biserial correlations (ρ_{eik}) and the difficulty values (p_{eik}) for each option of the test were first obtained. These biserial and difficulty values were then related to the IRT difficulty (or b) and discrimination (or a) parameters. The relationship existing between the biserial correlation and the item difficulty with the IRT parameters a and b is well established in the dichotomous case (Crocker & Algina, 1986; Lord, 1980; McDonald, 1985). In the polytomous case this relationship can also be used provided that the biserial correlations and the item difficulty values are computed on all the alternatives of the test. The relationship between the biserial correlations and the

item difficulty values with the a_{eik} and b_{eik} parameters is then as given:

$$a_{eik} = \frac{p_{eik}}{\sqrt{1-p_{eik}^2}}, \quad (41)$$

and

$$b_{eik} = \frac{\Phi^{-1}p_{eik}}{-p_{eik}}, \quad (42)$$

where

p_{eik} = the biserial correlation of alternative e to item i for group k;
 $\Phi^{-1}p_{eik}$ = the z-score that corresponds to the difficulty of the item (p) of alternative e to item i for group k.

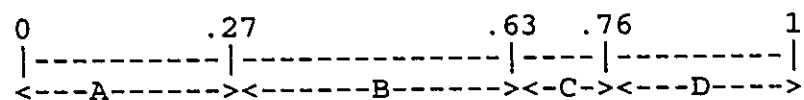
The items that were used as the basis for the input parameters were chosen from an actual test in the following way. Twenty-five items were selected from an actual test so that the difficulty of the items was similar to a uniform distribution. The actual test used to provide these input parameters was a 9th grade geography test from the Ontario Assessment Instrument Pool that originally had 41 items. Based on 323 examinees, the test mean and standard deviation for the 25 chosen items were 14.78 and 4.38, respectively. This 25 item test, with the 323 examinees had a test reliability (Kuder-Richardson 20) of 0.74.

The other condition of test length was achieved by using the characteristics of the 25 items twice to obtain a 50 item test. This procedure was chosen in order to ensure that the difficulty of the tests did not vary among the conditions. Once the input parameters a_{eik} and b_{eik} were obtained, the item responses for each sample were generated. This procedure is explained in the next section.

Generating the item responses

The procedures that was followed to generate item responses for the two groups are explained in this section. The procedures used to obtain the input parameters a_{eik} and b_{eik} were described above. In order to obtain the probability of obtaining alternative e for item i , the ability (θ) of the subject needs to be known (see equation 40). Each subject's ability (θ) was randomly generated from a normal distribution ($N(0,1)$). In this study, the two groups had the same underlying ability distribution. The abilities were generated using the International Mathematics and Statistics Library (IMSL) subroutine GGNML. Based upon the ability (θ) value generated for the subject, the probability of responding to each alternative of item i was computed for subject j by including this value of θ into equation (40).

Once the probability of responding alternative e of item i was known for subject j , the next step consisted of establishing the answer to item i for subject j . The total probability of responding to item i ranges between 0 and 1. This range was divided into areas which corresponded to the probability of selecting each alternative of item i . For example, suppose item i had 4 alternatives and the probabilities of selecting alternatives A, F C, and D were 0.27, 0.36, 0.13, and 0.24, respectively for subject j . Then, the probability range can be divided as follows;



The alternatives were ordered so that they occupied a distinct area on the probability scale. The answer to item i for subject j was obtained by generating a random number between 0 and 1 from a uniform distribution using the IMSL subroutine GGUBS. The location of this random number on the probability scale indicated the answer obtained by subject j to item i . For example, if the random number was equal to 0.55, subject j would have selected option B. The correct alternative to that item was known because the items were simulated based on actual test items. The total test scores for each of the subjects were computed by summing the items which had been answered correctly.

Computer programs

The program used to generate the item responses is presented in this section. Once each data set was simulated the data were analyzed with the five statistical approaches using standard statistical software packages (BMDP, 1990; SAS, 1990; SPSS, 1990). The procedures used to compute the statistical approaches studied in this investigation are also presented in this section.

Item response program

The program used to generate the item responses is provided in Appendix B. This program generates item responses for one group at a time. Both groups were generated by running this program twice. The input parameters (difficulty and discrimination), number of items, sample size, and starting seed were changed accordingly to simulate the appropriate condition.

Computing the DS approach

To compute the DS approach the data set for both the reference and focal groups was first analyzed with SAS (1990) using the PROC CORRESP procedure. This computed the first and second dimension weights for the whole test. Note, that the first and second dimension weights were computed with the two groups combined. Then, as mentioned earlier, the item responses were then substituted with the

weights of the second dimension. Once this was done a traditional ANOVA, using SPSS (1990), was computed on the weights of the first item. For this analysis the dependent variable was the weights of the second dimension for the first item and the independent variable was the grouping variable. The ANOVA results provided a test of uniform PDIF. PDIF was tested at a .05 level of significance.

Computing the LOG approach

The LOG approach was computed with the LOGLINEAR procedure in SPSS (1990). The three designs were:

Model 1 [A], [G], [R], [AG], [AR], [GR], [AGR];

Model 2 [A], [G], [R], [AG], [AR], [GR];

Model 3 [A], [G], [R], [AG], [AR];

where

A	is the ability of the subject;
G	is the group into which the subject belongs;
R	is the alternative chosen by the subject.

Once these various models were computed, uniform PDIF was computed by computing the difference in the Likelihood ratio between models 2 and 3. The significance of this result was tested using a .05 level of significance, with the degrees of freedom equal to the difference in degrees of freedom between the two models. Nonuniform PDIF was similarly tested by taking the difference between model 1 and model 2.

Computing the MCA and MCO approaches

For the MCA and MCO approaches, the responses for the item were saved as a 1 and 0 indicator matrix, where a subject was assigned a 1 to the option chosen and a 0 to the other options. In Table 6, an example of an indicator matrix for one item is provided. The data, in Table 6, were reparametrized into J-1 new options by comparing the first alternative with the last, the second alternative with the last one, and finally the third alternative with the last one. Once this transformation was done the MCA and MCO approaches were computed with SPSS (1990).

Table 6: Example of an indicator matrix for item i

Subject	Alternatives to item i			
	1	2*	3	4
1	0	1	0	0
2	1	0	0	0
3	0	1	0	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
N	0	0	1	0

MCA was computed by using a traditional MANOVA procedure with the only exception being that a sequential procedure was requested. In the model, the ability effect was first tested, followed by the group effect, and

finally the group-by-ability effect was tested. This sequential procedure ensured that the sums of squares for each group were independent.

In contrast the MCO approach was computed with a traditional MANCOVA. The covariate in this case was the ability of the subject and the independent variable was the group of the subject.

For both approaches the results of the Pillai-Bartlett (Bartlett, 1939; Pillai, 1955) trace statistics were recorded. The .05 level of significance level was used for all tests.

Computing the PLR approach

The BMDP (1990) PR procedure was used to compute the PLR approach. The analysis was done by forcing the independent variables into the model. This was needed in order to test the significance for the ability, group, and group-by-ability effects. Also, the order into which these effects were forced into the model was important in order to get a group effect corrected and independent of any differences in ability. To test each effect a chi-square difference test between sequential models was used. Again, a .05 level of significance was used to assess if any significant improvement in chi-square resulted by adding the group effect or by adding the group-by-ability interaction.

Uniform and nonuniform test of hypothesis

Uniform and nonuniform DIF are usually defined in terms of a logit ($\ln(P_i(1-P_i))$), where P_i is the proportion of alternative e to item i . These definitions were also extended to PDIF (see Chapter 2). However, depending on the approach employed, the uniform and nonuniform test of hypotheses are not necessarily formulated in terms of logits, but instead are defined in terms of differences in proportions of choosing alternative e to item i .

Logit differences

For the Logistic (LOG and PLR) approaches, uniform and nonuniform PDIF are both defined in terms of logit differences. Under these approaches, some or all the alternatives are said to display uniform PDIF when there is a constant logit difference between the groups at all the ability levels. There are some situations where some or all the alternatives of an item show a logit difference at lower ability levels favouring one subgroup and at higher levels it displays a difference in logit favouring the other subgroup. In these cases PDIF is said to be nonuniform.

Differences in proportions

For the MANCOVA (MCA and MCO) approaches uniform and nonuniform PDIF are defined in terms of differences in

proportions between the groups. Uniform PDIF under these approaches occur when the differences in proportions, on some or all the alternatives, between the groups of interest are constant across all the ability levels. Nonuniform PDIF would occur if the differences between the proportions, on some or all the alternatives, for the lower ability levels favoured one group while at higher ability levels it favoured another group. It should be mentioned that uniform PDIF with the DS is also defined in terms of differences in proportions.

Type of PDIF simulated

These discrepancies in the definitions among the approaches are important to understand because they affect the interpretation of the results. Actually, the type of PDIF simulated in this study does not allow for a direct comparison of the nonuniform test of hypotheses between the Logistic and MAN(C)OVA approaches. The uniform test of hypotheses, although different between the MAN(C)OVA and Logistic approaches can be compared. In the case of the Logistic approaches, the hypothesis tested is that the average difference in logits between groups is zero. In the MAN(C)OVA approach a test of equality of average differences in proportion is made.

Varying only the b parameter can result in slight nonuniform differences for the MAN(C)OVA approaches

because changing only the probability of choosing the alternatives can result in both a difference in the proportion and a difference in the biserial correlation value. This occurs because the biserial correlation value of an alternative is essentially the correlation between the proportion of subjects who choose that alternative with the subjects total test scores. Thus, it becomes obvious that a change in the probability of choosing alternative e has an effect on the biserial correlation value of alternative e .

The difference in PDIF simulated between the Logistics and MAN(C)OVA approaches may be better explained graphically. This is done by showing the characteristic curves for the 0.70 PDIF conditions for each of the three contrasts tested by the approaches. Note, that the three contrasts represent the differences of each the first, second, and third alternatives with the last alternative of the item. These characteristic curves are shown for both the PDIF 1 and 2 conditions.

PDIF 1 conditions. The characteristic curves for each of the PDIF 1 contrasts are shown in Figures 1, 2 and 3. These curves show the differences simulated between the reference and focal groups in terms of proportions. These Figures clearly indicate that changing the probability of the alternatives by varying the b does not yield only a

uniform difference in proportions but it also introduces a nonuniform difference in proportions. This, actually, occurs for all the contrasts.

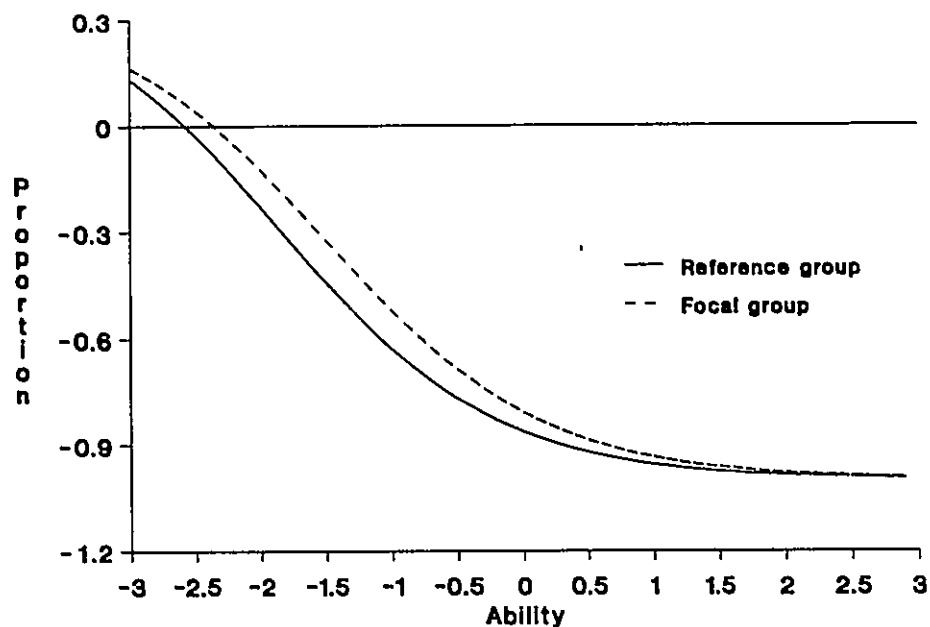


Figure 1: Contrast 1 characteristic curves for the PDIF 1 condition (proportion)

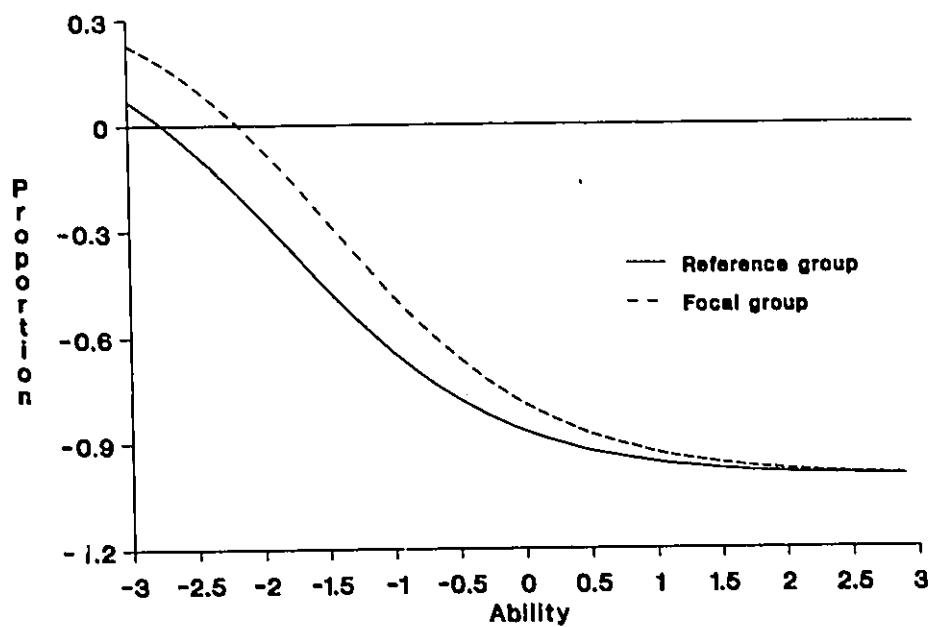


Figure 2: Contrast 2 characteristic curves for the PDIF 1 condition (proportion)

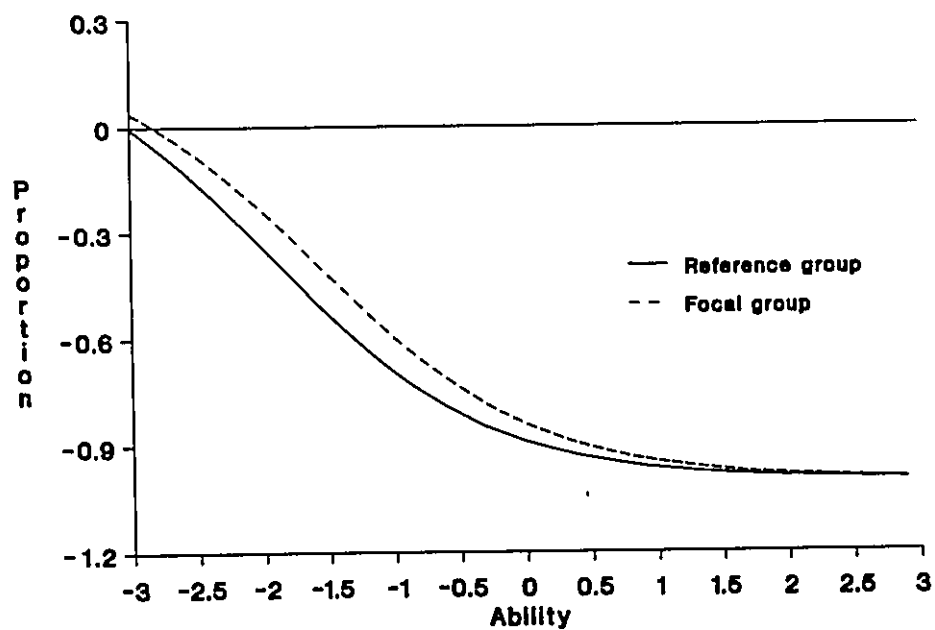


Figure 3: Contrast 3 characteristic curves for the PDIF 1 condition (proportion)

In contrast, changing the probability of choosing the alternatives by altering the b value does correspond to a uniform logit difference between the reference and focal groups. This is observed from Figures 4, 5, and 6. For each of the contrasts in the PDIF 1 conditions the difference simulated between the reference and focal groups is almost constant across the ability scale.

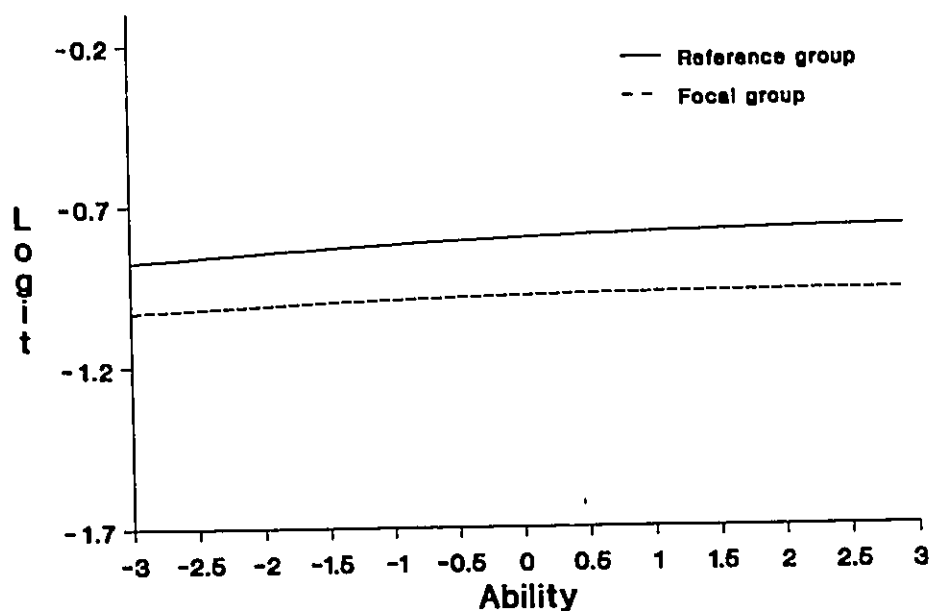


Figure 4: Contrast 1 characteristic curves for the PDIF 1 condition (logit)

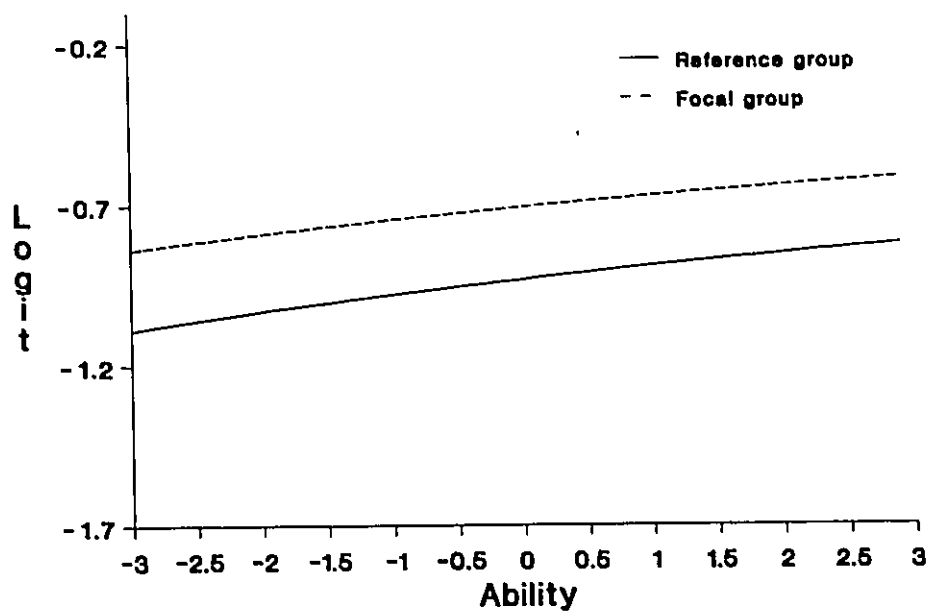


Figure 5: Contrast 2 characteristic curves for the PDIF 1 condition (logit)

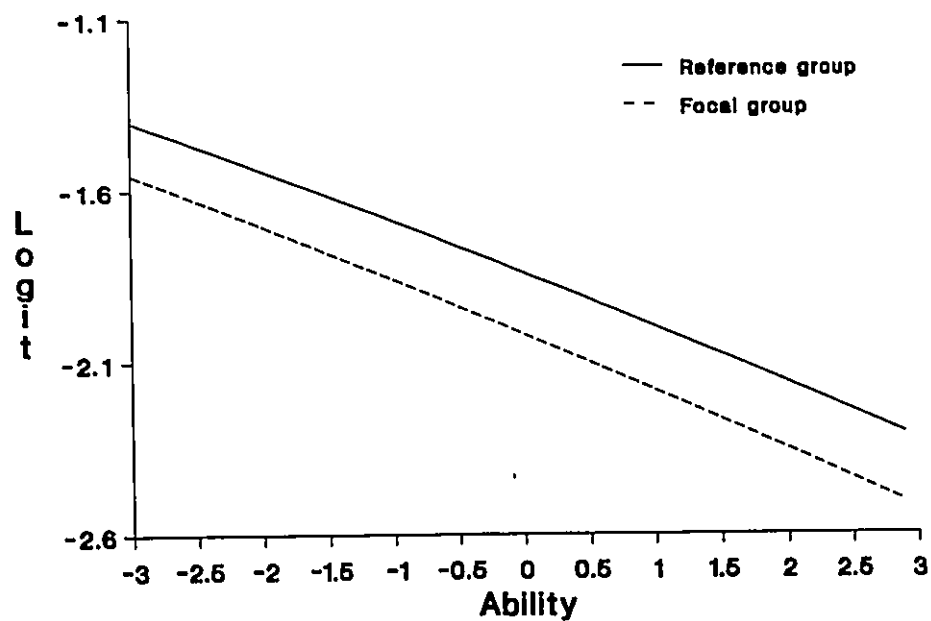


Figure 6: Contrast 3 characteristic curves for the PDIF 1 condition (logit)

From these Figures, it becomes obvious that the type of PDIF simulated is not directly comparable between the Logistic approaches and those of the MAN(C)OVA and DS approaches. The MAN(C)OVA approaches are expected to detect a significant uniform difference in proportions; however, they are also expected to detect some nonuniform differences in proportions. These results are expected when PDIF has been introduced in the alternatives of the item because as observed in Figures 1, 2, and 3 the PDIF simulated did not yield only uniform differences in proportions. In contrast, the Logistic approaches are only expected to detect a uniform logit difference as the PDIF simulated did result in creating only uniform logit differences between the groups.

PDIF 2 conditions. The same observations can be made with the 0.70 PDIF 2 conditions. Again, the differences in proportions simulated between the reference and focal groups show that the PDIF is not uniform across the ability scale. This difference is shown for each of the contrast in Figures 7, 8, and 9. Also, the logit differences simulated are observed to be almost constant across the ability scale (Figures, 10, 11, and 12).

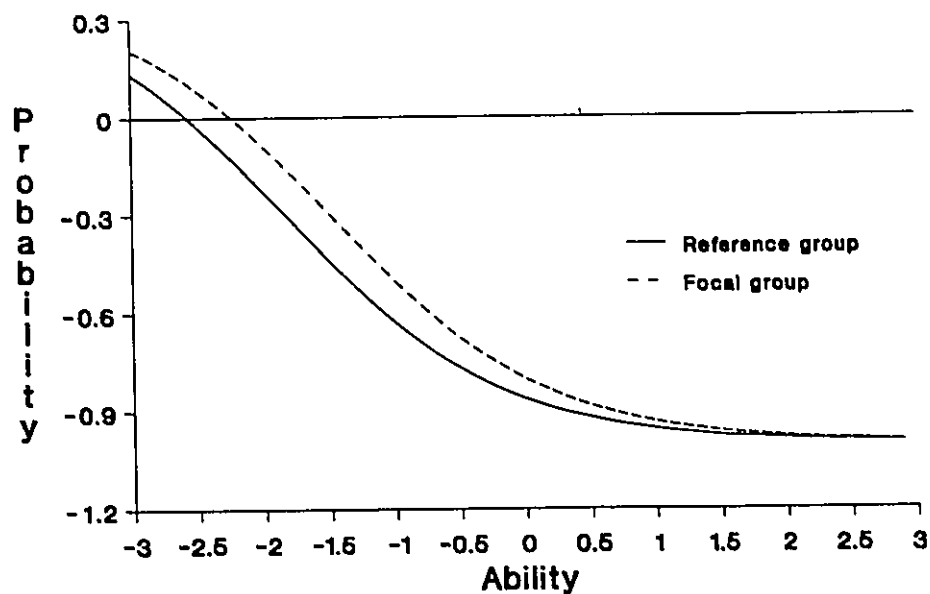


Figure 7: Contrast 1 characteristic, curves for the PDIF 2 condition (proportion)

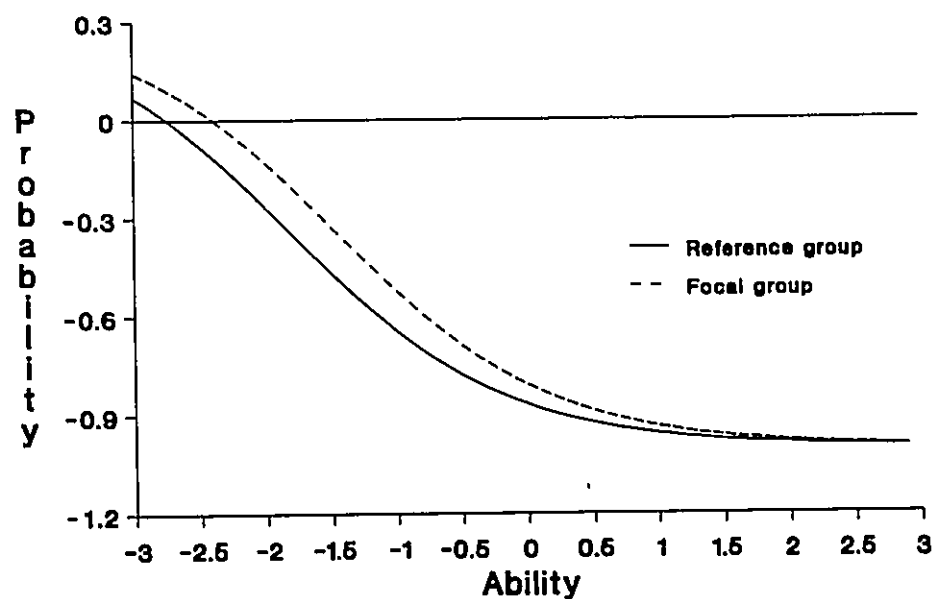


Figure 8: Contrast 2 characteristic curves for the PDIF 2 condition (proportion)

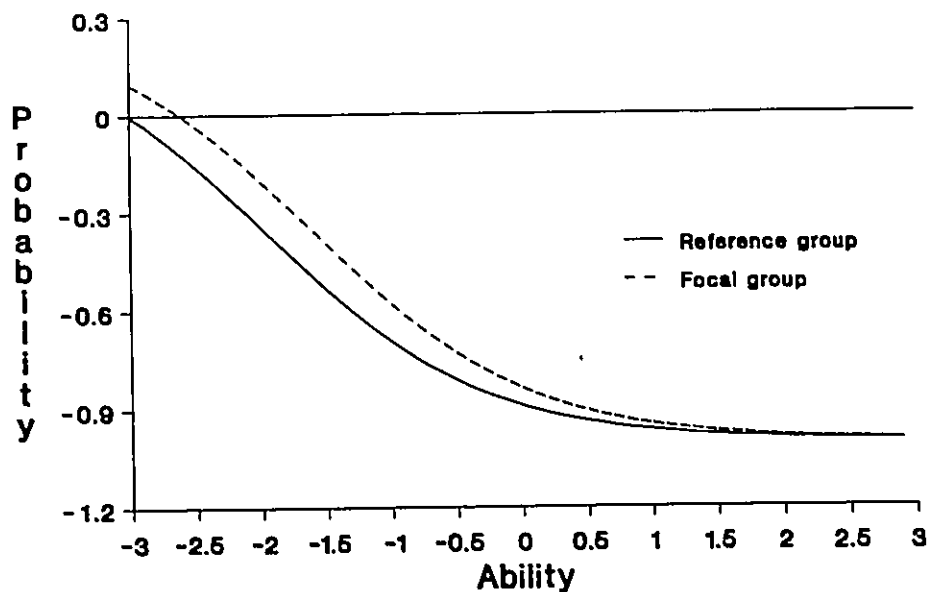


Figure 9: Contrast 3 characteristic curves for the PDIF 2 condition (proportion)

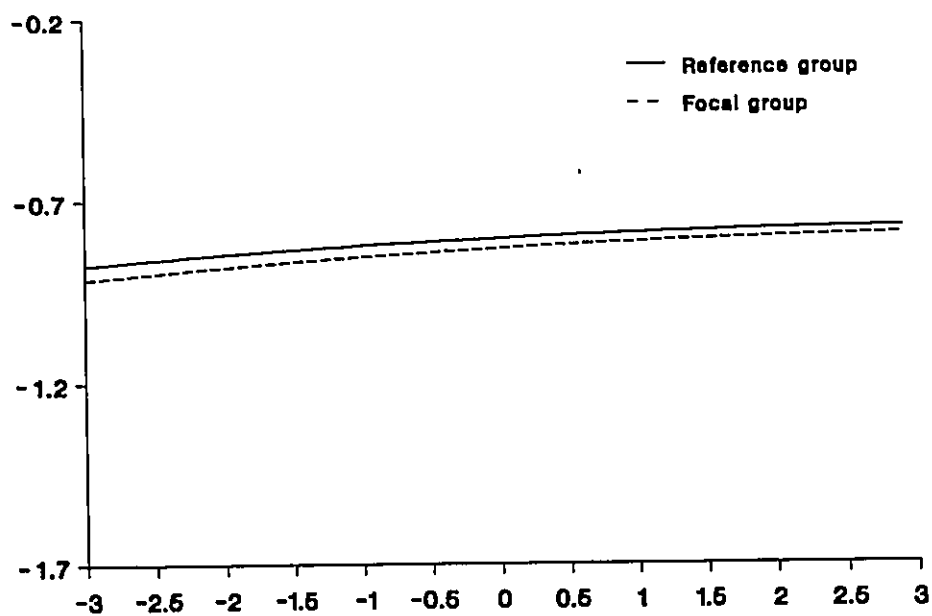


Figure 10: Contrast 1 characteristic curves for the PDIF 2 condition (logit)

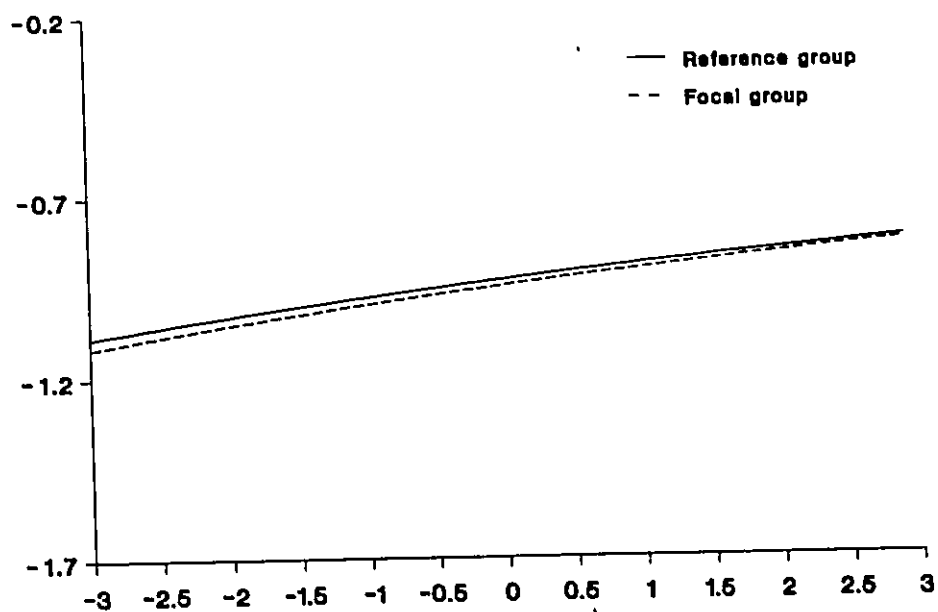


Figure 11: Contrast 2 characteristic curves for the PDIF 2 condition (logit)

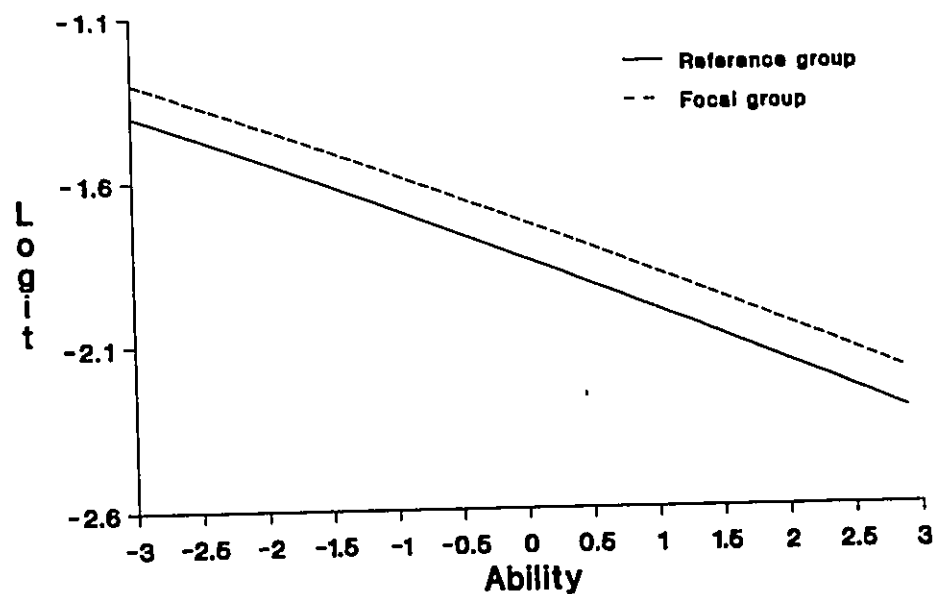


Figure 12: Contrast 3 characteristic curves for the PDIF 2 condition (logit)

Data analysis

The Type I error rates at the .05 level of significance were reported for the NO PDIF and PDIF conditions. The rejection rates were compared to the theoretical values. The theoretical values have a confidence interval which is plus or minus 1.965 standard errors of a proportion (S.E.P.) from the expected nominal value. It should be mentioned that the plus or minus 1.965 value was chosen because it corresponds to a 95% confidence interval. The theoretical standard error of a proportion is

$$S.E.P. = \sqrt{\frac{p(1-p)}{N}}, \quad (57)$$

where

p = is the expected nominal value;
N = is the number of replications.

Any empirical rejection rates that were greater than 1.965 standard errors of a proportion from the expected nominal values were considered significantly different.

To investigate the effect of test length, sample size, and item difficulty on the Type I error rates a logit-linear analysis was conducted. Similarly, to investigate the effect of test length, sample size, item difficulty, and the amount and location of PDIF on the

rejection rates, when PDIF was introduced, a logit-linear analysis was conducted.

Research questions

This study investigated the following research questions:

- (1) What are the Type I error rates associated with the DS, LOG, MCA, MCO, and the PLR approaches?
- (2) How powerful are the DS, LOG, MCA, MCO, and the PLR approaches in detecting uniform PDIF?
- (3) Do test length, sample size, item difficulty, amount and location of PDIF have any effect on the rejection rates of the DS, LOG, MCA, MCO, and the PLR approaches?

Limitations of the study

Some of the limitations within which this investigation was conducted can be summarized as follows:

- (1) Conducting a Monte-Carlo investigation limits the extent of generalization of the results to the conditions that were examined in this study.

- (2) The two-parameter polytomous logistic regression model employed to generate the data might yield some unrealistic test results because the lower asymptote was set to zero. In a real testing situation, the lower ability examinees might be expected to correctly guess the answer of some difficult items. Omitting this factor in the model is expected to yield higher test reliability than would occur with actual item response data.
- (3) The underlying ability of the subjects for the reference and focal groups was set to be equal for all the conditions investigated in this study. The results of this study are, therefore, limited to equal abilities between groups. Also, the underlying ability was assumed to be normally distributed. This is a common assumption that yields results that are generalizable only to the extent to which this is true.
- (4) IRD in this study are analyzed at the item level only. Some authors would suggest that if there are no test response differences there is no need to further look at IRD. This study is based on the premise that IRD procedures are especially needed

when developing a test in order to prevent eventual test response differences.

- (5) This study was limited to the analysis of PDIF that was introduced in only one item of the test. In the future, it would also be interesting to investigate the effect of having several items exhibiting PDIF in the test.
- (6) Only one specific type of PDIF, uniform PDIF, was introduced in this study. Actually, the b or difficulty parameter was the only parameter varied in this study. This study, therefore, does not provide any evidence on the effectiveness of these approaches when only the a or discrimination parameter is varied or when both parameters are varied.
- (7) The number of replications, in this study, for each condition was set at 500 replications. Researchers might consider this to be too small to yield precise results. However, it is considered to be adequate to get a general idea of the performance of the approaches.
- (8) In this study, the difference in sample size between the focal and reference groups was set at a one-to-

two ratio. It is acknowledged that in many testing situations that ratio might differ.

CHAPTER 5

Results

In this section, the results of this investigation are presented, as well as an example of the results obtained with the data generation program. The latter is provided to show that proper data generation procedures were followed. The results for the uniform tests of hypothesis are presented first for the NO PDIF conditions and then for the PDIF conditions. Next, the results for the nonuniform tests of hypothesis are presented for the NO PDIF conditions and the PDIF conditions.

The effects of the independent variables on the statistical decisions (i.e. the numbers of acceptances and rejections) were analyzed using logit-linear analyses. A logit-linear analysis is a log-linear analysis with one variable specified as the dependent variable. The main effect of any independent variable, with this analysis, actually represents an interaction with the dependent variable (Baker, 1981). Bock's approach (Bock, 1975, chap. 8) was followed to obtain the most parsimonious logit-linear model. This approach used a nested hierarchical model, meaning that all the lower associations nested within the model were included. Also, parsimony under Bock's approach was obtained by building the model in a forward stepwise manner. Forward model-building starts with simple main effects and more complex associations

were added until a hierarchical model was found to be significant. The overall evaluation of the model was made using the likelihood ratio chi-square statistic. A model with a p -value greater or equal to 0.15 was considered adequate. The standardized parameter estimates were used to assess the significance of individual effects. These parameters were interpreted as a standard normal deviate (z). Therefore, in order to assess their significance they were compared with a critical z . The standardized parameter estimates were considered significant if their absolute values were greater than 1.965. The independent variables for the logit-linear analyses were test length, sample size, item difficulty, and the location and amount of PDIF. The effects of the independent variables on the statistical decision were analyzed separately for each of the statistics under investigation, namely, the Dual Scaling (DS), Log-Linear (LOG), Categorical MANOVA (MCA), Continuous MANOVA (MCO), and Polytomous Logistic Regression (PLR) approaches. For simplicity, the results are presented in terms of the independent variable(s) only. However, in a logit-linear analysis this implies an interaction with the dependent variable (i.e. statistical decision). For example, if a sample size-by-statistical decision was significant, it was referred to only as a sample size main effect.

The numbers of Type I errors results were compared with their theoretical values by using the standard error of a proportion $\{p(1-p)/N\}^{1/2}$, where p is the expected nominal value of significant effects and N is the number of replications. The theoretical α value has a 95% confidence interval which is plus or minus 1.965 standard errors of a proportion. This confidence interval based on 500 replications is $0.0305064 \leq \alpha \leq 0.0694935$. This confidence interval can be expressed in terms of numbers of rejections expected in 500 replications, where the number of expected rejections is 25 ($N*\alpha$). Therefore, this theoretical number of rejections has a confidence interval bounded by 15.25 and 34.75. This confidence interval was the one used to assess the observed numbers of Type I errors.

Data generation verification

The purpose of this section is to provide some evidence that the data were properly generated. This is accomplished by providing an example of the results which are obtained with the data generation program. The first section is concerned with providing some evidence that the two IMSL subroutines used by the data generation program were functioning as they should have been. The second section is concerned with providing some evidence that proper results were obtained with the data generation program.

Verification of the IMSL subroutines

The underlying abilities of the subjects in each group were assumed to be normally distributed. In order to generate such a distribution the IMSL subroutine GGNML was used. The accuracy of this subroutine to generate a normal distribution of ability with a given mean and standard deviation is shown later in this section. The GGUBS subroutine was used to determine the alternative to the item the subject chose. Once the probabilities of choosing the alternatives of the item were calculated for the subject a criteria was used to determine which alternative the subject chose. This was done by ordering the subject's probabilities of choosing each of the alternatives on the probability scale. Once this was done a uniformly distributed random number was generated, using the GGUBS subroutine, to ascertain the answer of that subject on the item. For further explanation refer to the methodology section. This section provides some evidence that the numbers generated with the GGUBS subroutine were uniformly distributed.

The procedures followed to assess the proper functioning of both the GGNML and GGUBS subroutines consisted of calling each subroutine 1000 times and recording the generated values. These results are presented in the next sections.

GGNML subroutine. The results of the 1000 calls to the GGNML subroutine are summarized in Table 7. It is observed from Table 7 that the mean and the standard deviation values are very close to the expected mean and standard deviation of a normal standard deviate distribution ($N(0,1)$). The other indices also indicate that the standard deviates generated were normally distributed (i.e. both the skewness and kurtosis values are close to zero). Figure 13, further confirms that the abilities generated were normally distributed as well as providing some evidence that most of the generated values fell within plus or minus three standard deviations. From these results it can be concluded that the underlying abilities of the group, generated with the GGNML subroutine, were in fact normally distributed.

Table 7: Summary of GGNML verification

Mean	0.005
Standard deviation	0.949
Minimum	-2.657
Maximum	3.471
Skewness	0.094
Kurtosis	0.144

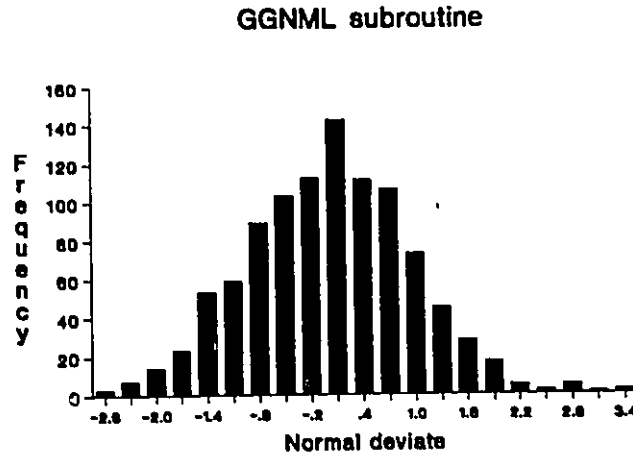


Figure 13: GGNML standard deviates distribution

GGUBS subroutine. The results of 1000 calls to the GGUBS subroutine are summarized in Table 8, and a plot of the random numbers generated by this subroutine are shown in Figure 14. Note that the midpoints of the intervals are displayed in Figure 14. These results indicate that the GGUBS subroutine generated random numbers between 0 and 1. Table 8 and Figure 14 clearly show that uniformly distributed random numbers were generated with the GGUBS subroutine.

Table 8: Summary of GGUBS verification

Mean	0.503
Standard deviation	0.287
Minimum	0
Maximum	1
Skewness	-.057
Kurtosis	-1.186

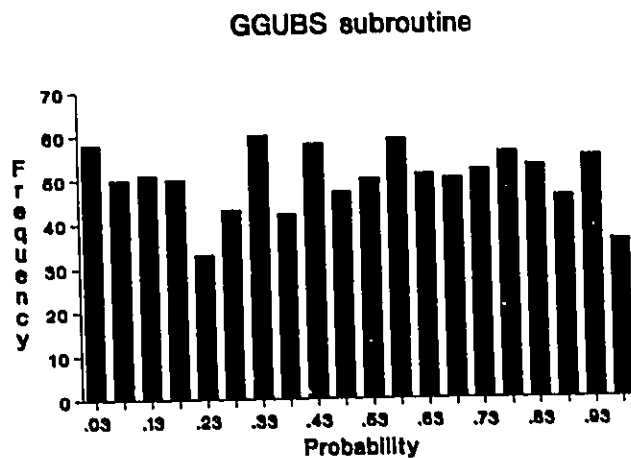


Figure 14: Distribution of the random numbers

Verification of the data generation program

Evidence of proper data generation is achieved by reporting the results of 100 replications for one of the conditions studied in this investigation. The condition presented was the NO PDIF condition with short item test (25 items), and with a small sample size (i.e. 500 subjects in the reference group and 250 in the focal group). This condition was chosen for two reasons. First, the smaller sample size conditions may yield less accurate results than the larger conditions, and second, the shorter test conditions may result in more differences in ability estimates than the longer test ones. Thus, the condition chosen is the one that may yield more differences in terms of replicating the input parameters and also it may yield more variation in the total test

score results than any other conditions. Finally, the NO PDIF condition was chosen because it made it possible to directly compare the total test score results of the reference group with those of the focal group. The choice of this condition is important because it allows for some verification of proper data generation because the test score results and the output parameters between groups are expected to be equal.

These verification results are presented in three main sections. The first section presents some descriptive results for the total test scores that were generated for both the reference and focal groups. These results are also compared with the basis group that was used to initially compute the input parameters used by the data generation program (further information pertaining to the basis group is provided in the methodology section). The second section provides some evidence that the four ability groups computed by the program were in fact comprised of approximately 25% of the subjects. These ability groups were derived by the program because some of the statistics used these ability groups as a measure of ability instead of the total test score. The results of the ability groupings are shown for both the reference and focal groups. Finally, the last section is concerned with providing some evidence that proper item parameters were generated with the data generation program. These

parameters are also contrasted with the ones that were initially used to generate them.

Total test score descriptive results. The results of the 100 replications for the NO PDIF condition with the 25 item test and small sample sizes are presented in Table 9, for both the reference and focal groups. Descriptive results for the basis group are also shown in Table 9. Table 9 displays the mean results of the 100 replications for the reference and focal groups with their standard deviation written below in parentheses.

Table 9: Descriptive statistics for the basis, reference, and focal groups

	Groups		
	Basis (n=323)	Reference (n=500)	Focal (n=250)
Mean	14.78	13.94 (.24)	13.90 (.36)
Standard Deviation	4.39	5.54 (.14)	5.57 (.17)
Skewness	-.234	-.336 (.063)	-.327 (.095)
Kurtosis	-.616	-.806 (.095)	-.710 (.043)
Reliability (KR-20)*	.747	.849 (.010)	.850 (.012)

* Kuder Richardson formula 20

From Table 9, it is quite clear that there are no differences between the results of the reference and focal

groups. The average results for the mean, standard deviation, reliability, skewness, and kurtosis were all found to be very similar for these two groups. No differences between these groups were expected because no PDIF was introduced in the alternatives of the test. These results provide some evidence that valid between-group comparisons can be made.

The results of both the reference or focal groups, however, differ somewhat from the basis group. This was expected because the underlying ability distribution used to generate the focal and reference groups were assumed to be normally distributed. However, the ability distribution of the basis group may have been different from normal. Differences in test distribution may easily explain the differences in total test score, standard deviation, skewness, kurtosis, and reliability. Also, it is well known that the reliability (KR-20) coefficient depends on the variability found in the test. Greater variability in the total test scores tends to increase the value of the reliability coefficient (Crocker & Algina, 1986). Thus, it is not surprising to observe greater reliability coefficients with the simulated groups (i.e. reference and focal groups) than the basis group, because greater variability was observed with these groups. Actually, it is possible to estimate the reliability of the basis group with a larger standard deviation by using the Kuder-

Richardson 21 formula (Crocker & Algina, 1986). Adjusting the reliability value of the basis group using the reference group standard deviation results yields a reliability value of 0.836. This adjusted reliability value then becomes quite comparable to the reference and focal groups.

Ability group distributions. Evidence that the data generation program classified the total test scores into four ability groups with the objective of having 25% of the subjects in each group is shown in this section. The mean results of the 100 replications for the NO PDIF condition with the 25 item test and the small sample are shown in Table 10. As can be seen in Table 10, this classification of subjects appears to have been accomplished successfully. Both groups had approximately 25% of the subjects in each ability group.

Table 10: Ability group distributions for 100 replications

Group	Ability group	Mean	Standard deviation
Reference	1	23.86%	1.81%
	2	25.70%	2.07%
	3	25.51%	1.75%
	4	24.92%	2.24%
Focal	1	24.14%	2.60%
	2	25.68%	2.88%
	3	25.49%	2.48%
	4	24.68%	3.01%

Input and output parameters. In this section results relating to the estimation of the input parameters (used to generate the item responses) are presented. The means of the difficulty and biserial parameters based on the 100 replications of the reference and focal groups are found in Appendices C and D, respectively. The actual difficulty and biserial correlation parameters of the basis group are also shown in these appendices. In order to assess the relationships among the groups on the difficulty and biserial correlation parameters linear regression equations were computed. A summary of these results are presented in Table 11.

Table 11: Summary of the regression analyses among the reference (R), focal (F), and basis (B) groups

Prediction parameter	Eq. number	Equation	Correlation
difficulty	1	$B = -.065^* + 1.170^*R$	.991
	2	$B = -.062^* + 1.170^*F$	.992
	3	$R = -.003 + .998^*F$	1.000
biserial	4	$B = -.047^* + 0.812^*R$	.963
	5	$B = -.046^* + 0.806^*F$	.961
	6	$R = -.001 + 0.995^*F$	.999

The asterisk denotes the coefficients which are significant at an alpha of 0.05.

In Appendix C, it is observed that the difficulty parameters for the reference and focal groups are similar

to the basis group. Actually, as shown in Table 11, the difficulty parameters are highly correlated with those of the basis group ($r=.991$ and $r=.992$ respectively). Regression equations 1 and 2, however, indicate that the difficulty parameters of the basis group are 1.17 times larger than those of the reference and focal groups. In contrast, the difficulty parameters of the reference and focal groups were found to be very similar (see Appendix C). This observation is further confirmed by equation 3 of Table 11. A near perfect relationship was found between the difficulty parameters of the reference and focal groups.

The comparison of the biserial correlation parameters between the basis group with those of the reference and focal groups were similar to those of the difficulty parameters (see equations 4 and 5 of Table 11, and Appendix D). Again, high correlations were found between the reference ($r=.963$) or focal groups ($r=.961$) with the basis group. However, a factor of approximately 0.81 was needed to get the basis group biserial correlation parameters from either the reference and focal groups. Similar to the difficulty parameters, the biserial correlations were almost identical between the reference and focal groups (see Appendix D). This is further confirmed with regression equation 6 in Table 11. The

biserial correlations for the reference and focal groups are almost perfectly correlated.

In summary, these results indicate that the reference and focal groups' difficulty and biserial correlation parameters were different than the basis group. These slight differences in the difficulty and biserial correlation parameters between the simulated groups with those of the basis group were somewhat expected. This may have occurred because the underlying distribution of ability for the generated groups was assumed to be normally distributed but for the basis group the abilities may have had another distributional shape. Also, some of those differences might be related to the model used to generate the data (i.e., a two-parameter polytomous logistic regression model). Perhaps, the simulated data sets might resemble better the real data set if guessing was accounted for in the model.

Finally, the comparison of the difficulty and biserial correlation parameters between the reference and focal groups were found to be nearly identical. This indicated that valid between-group comparison can be made.

Test of uniform PDIF

The results pertaining to the tests of uniform PDIF are presented in this section. The test of uniform PDIF, as mentioned earlier, checks the groups under investigation to see if they have different patterns of responses on the studied item. The results for the NO PDIF conditions are presented first. These results are followed by those of the PDIF conditions. Because no group differences were introduced in the NO PDIF conditions, the observed number of rejections for these conditions correspond to the number of Type I errors. In contrast, for the PDIF conditions uniform differences were introduced in the alternatives of the item between the groups and therefore, these tests of hypothesis assess the power of each statistical approach in its effectiveness to detect uniform differences in the alternatives of the item (i.e. test of uniform PDIF).

NO PDIF conditions

The numbers of rejections for the NO PDIF uniform tests of hypotheses based on 500 replications are shown in Table 12. The results indicate that the numbers of Type I errors for the MCO and the PLR approaches were within the expected theoretical numbers of rejections in all the conditions. The numbers of Type I errors for the DS, LOG, and, MCA approaches were above the expected values in some

conditions. All these approaches had elevated Type I error rates for the 0.70 difficulty condition with 50 items and with small sample sizes (500/250). The number of DS rejections for the 0.40 difficulty condition with 25 items and the larger sample size (1500/750) was slightly below the expected theoretical number of rejections.

Table 12: Uniform tests of hypothesis for the NO PDIF conditions

Item Difficulty	Test Length	Sample Size	Statistics				
			DS	LOG	PLR	MCA	MCO
p=.70	25 items	500/250	18	24	25	25	26
		1500/750	21	23	25	22	25
	50 items	500/250	36	36	32	36	32
		1500/750	31	28	28	30	28
p=.40	25 items	500/250	27	29	25	28	26
		1500/750	14	31	33	30	34
	50 items	500/250	22	29	22	30	24
		1500/750	16	27	24	28	25

Numbers in bold are more/less than 1.96 standard errors from the expected number (25) of rejections.

The results of the separate logit-linear analyses indicate that for the LOG ($\chi^2(7)=4.27$, $p=.748$), MCA, ($\chi^2(7)=4.38$, $p=.736$), MCO ($\chi^2(7)=3.43$, $p=.842$), and PLR ($\chi^2(7)=4.15$, $p=.762$) approaches, a model with only the dependent variable (i.e. statistical decision) was sufficient to explain the distribution of frequencies observed. Therefore, none of the independent variables had

any effect on the Type I error rates of these approaches. However, for the DS approach a model incorporating the interaction of test length by item difficulty as well as the main effects of test length and item difficulty was needed to sufficiently explain these results ($\chi^2(4)=6.00$, $p=.199$). This interaction can be explained in averaging the frequencies across the independent variables not included in the model (see Table 13). It appears, from Table 13, that the long test length conditions with the easier item ($p=.70$) had a larger number of rejections than any other conditions.

Table 13: DS average number of Type I errors for the test length by item difficulty interaction

Item Difficulty	Test length	
	25 items	50 items
$p=.70$	19.5	33.5
$p=.40$	20.5	19

PDIF conditions

This section describes the results obtained for the tests of uniform PDIF, when a uniform group difference was introduced into the alternatives of the item. These empirical numbers of rejections actually depict the power of each statistical approach.

The numbers of rejections obtained for the 0.70 and 0.40 PDIF conditions are shown in Table 14 and 15,

respectively. Similar results were obtained for the LOG, MCA, MCO, and PLR approaches. A logit-linear analysis was conducted for each of the four approaches separately in order to determine those factors that affected the statistical decision. The results of the logit-linear analyses revealed that the same model described the numbers of rejections for all four approaches. Main effects of sample size, item difficulty, and location of PDIF were found to significantly influence the probability of rejecting the null hypothesis for the LOG ($\chi^2(12)=11.88$, $p=.456$), MCA ($\chi^2(12)=12.15$, $p=.434$), MCO ($\chi^2(12)=11.74$, $p=.466$), and PLR ($\chi^2(12)=12.89$, $p=.377$) approaches. The effects of sample size, item difficulty, and location of PDIF are quite clear. It appears that larger sample sizes increase the number of rejections, whereas the more difficult item conditions were found to have a lower number of rejections. Finally, larger numbers of rejections were observed with the PDIF 1 condition.

Table 14: Numbers of rejections for the 0.70 PDIF conditions

PDIF Condition	Test Length	Sample Size	Statistics				
			DS	LOG	PLR	MCA	MCO
1	25 items	500/250	71	116	120	121	122
		1500/750	110	320	310	323	312
	50 items	500/250	67	134	120	140	124
		1500/750	124	316	318	320	324
2	25 items	500/250	40	94	92	101	100
		1500/750	58	253	258	257	256
	50 items	500/250	43	89	89	98	94
		1500/750	72	247	247	250	246

Table 15: Numbers of rejections for the 0.40 PDIF conditions

PDIF Condition	Test Length	Sample Size	Statistics				
			DS	LOG	PLR	MCA	MCO
1	25 items	500/250	50	86	87	83	92
		1500/750	114	229	233	231	238
	50 items	500/250	67	95	94	93	97
		1500/750	145	247	254	250	248
2	25 items	500/250	35	60	57	62	59
		1500/750	47	163	168	163	164
	50 items	500/250	36	60	71	62	63
		1500/750	82	193	195	194	191

The results showed that for each condition investigated, the DS approach had much lower numbers of rejections than all the other approaches. The results of the logit-linear analysis were slightly different than

those for the other statistical approaches. A model including the main effects of test length, sample size, and the location of PDIF was necessary to explain the observed frequencies ($\chi^2(11)=14.44$, $p=.210$). Similar to the other approaches, the larger sample size conditions were found to have a larger number of rejections than the smaller sample size conditions. As well, larger numbers of rejections were observed for the PDIF 1 conditions. Also, larger numbers of rejections were observed with the longer test length conditions.

Test of nonuniform PDIF

The results of the tests of the nonuniform hypothesis are presented in this section. These results are first presented for the NO PDIF conditions and then for the PDIF conditions.

In this investigation only uniform logit differences were introduced. Therefore, the nonuniform tests of hypothesis results should provide a test of a true null hypothesis. For the NO PDIF conditions all the approaches provided a test of a true null hypothesis. However, this was not true for the PDIF conditions. This, as was mentioned in the methodology section, occurs because some of the approaches (LOG and PLR) test for nonuniform logit differences, while others test (MCA and MCO) for nonuniform differences in proportions. The introduction of

logit uniform differences translates into a uniform difference in proportions as well as slight nonuniform differences in proportions. The observed PDIF results for the MCA and MCO approaches were adjusted to allow for an approximate comparison with the LOG and PLR results. The observed and adjusted results are presented in the PDIF section along with the procedure used to compute these Adjusted Numbers of Rejections (ANR).

NO PDIF conditions

The results of the nonuniform tests of hypothesis, when no differences between the groups were introduced, are presented in this section. These results correspond to a test of the null hypothesis for all the approaches. The numbers of rejections for the NO PDIF nonuniform tests of hypotheses observed in 500 replications are shown in Table 16. These results indicate that each statistical procedure had conditions in which the numbers of Type I errors were significantly different from the expected nominal values.

Table 16: Nonuniform tests of hypothesis for the NO PDIF conditions

Item Difficulty	Test Length	Sample Size	Statistic			
			LOG	PLR	MCA	MCO
p=.70	25 items	500/250	35	24	34	36
		1500/750	15	25	26	48
	50 items	500/250	39	20	26	37
		1500/750	25	24	35	39
p=.40	25 items	500/250	36	18	28	17
		1500/750	21	27	19	18
	50 items	500/250	37	35	28	16
		1500/750	29	35	31	24

Numbers in bold are more/less than 1.96 standard errors of the expected number (25) of rejections.

In some conditions the numbers of Type I errors for the LOG approach were found to be significantly above the expected number, while in other conditions fewer rejections were found than would have been expected. The numbers of rejections for the LOG conditions were found to be slightly higher than expected for all the small sample size conditions, while for the large sample size conditions they tended to be within the nominal confidence limits. The exception is for the 25-item test condition with the 0.70 difficulty level and with the large sample size condition. Here the actual number of rejections is slightly lower than expected. The results of the logit-linear analysis indicate that, for the LOG approach, a model with sample size as the sole independent variable

adequately explains the statistical decision variations ($\chi^2(6)=5.41$, $p=.493$). It can be observed from Table 16 that smaller numbers of rejections were found in the larger sample size conditions.

The numbers of Type I errors for the PLR approach were slightly above the expected number of rejections for the 50-item test condition with a 0.40 difficulty value. In all the other conditions the empirical numbers of Type I errors fell within the nominal confidence limits. Logit-linear analysis results, for the PLR approach, indicate that a model with an interaction between test length and item difficulty as well as the main effects of test length and item difficulty sufficiently explain the variation observed in the dependent variable ($\chi^2(4)=2.30$, $p=.681$). In averaging the frequencies across the independent variables not included in the model, the average number of rejections for the longer test length with the 0.40 PDIF conditions was larger than for any other conditions (see Table 17).

Table 17: PLR average number of Type I errors for the test length by item difficulty interaction

Item Difficulty	Test length	
	25 items	50 items
p=.70	24.5	22
p=.40	22.5	35

The numbers of Type I errors for the MCA approach appeared to be inflated only in the conditions having a 50-item test and a large sample size. The logit-linear analysis indicates that a model with an interaction term between test length and sample size as well as the main effects of test length and sample size adequately explain the variation observed in the results ($\chi^2(4)=2.10$, $p=.717$). It was found, in averaging the frequencies of rejections across the independent variables not included in the model, that the average number of rejections of the null hypothesis was larger for the shorter test length in the small sample size conditions. As well, the average number of rejections of the null hypothesis was larger for the longer test length in the large sample size conditions (see Table 18).

Table 18: MCA average number of Type I errors for the test length by sample size interaction

Sample Size	Test length	
	25 items	50 items
500/250	31	27
1500/750	22.5	31

Finally, the numbers of rejections of the MCO approach were all within the expected confidence interval for the 0.40 item difficulty conditions. However, for the 0.70 item difficulty conditions, the numbers of rejections for the MCO test of nonuniform PDIF were all above the

expected nominal values. Results from the logit-linear analysis indicate that a model with item difficulty as the sole independent variable adequately explains the statistical decision variations ($\chi^2(6)=4.42$, $p=.620$).

PDIF conditions

As was mentioned earlier, only uniform logit differences were introduced in the alternatives of the item (i.e. uniform PDIF). Therefore, the number of rejections with the tests of nonuniform hypothesis should be consistent with expected number of Type I errors when the tests are carried out with the two LOGISTIC approaches (LOG and PLR). However, the MAN(C)OVA (MCA and MCO) approaches were expected to have a number of rejections above the nominal values because introducing uniform logit differences translates into both a uniform difference in proportions and a slight nonuniform difference in proportions.

The numbers of rejections for the tests of nonuniform PDIF with the 0.70 and the 0.40 item difficulty conditions are shown in Tables 19 and 20, respectively.

Table 19: Nonuniform tests of hypothesis for the 0.70 PDIF conditions

PDIF Condition	Test Length	Sample Size	Statistic			
			LOG	PLR	MCA	MCO
1	25 items	500/250	36	25	42	51
		1500/750	32	19	77	82
	50 items	500/250	47	23	47	49
		1500/750	25	30	44	75
2	25 items	500/250	33	23	42	36
		1500/750	32	22	61	50
	50 items	500/250	44	23	40	57
		1500/750	30	21	41	65

Numbers in bold are more than 1.96 standard errors from the expected number (25) of rejections.

Table 20: Nonuniform tests of hypothesis for the 0.40 PDIF conditions

PDIF Condition	Test Length	Sample Size	Statistic			
			LOG	PLR	MCA	MCO
1	25 items	500/250	33	27	52	20
		1500/750	27	22	53	33
	50 items	500/250	41	28	33	31
		1500/750	46	32	64	50
2	25 items	500/250	35	36	34	25
		1500/750	27	31	48	28
	50 items	500/250	35	24	36	21
		1500/750	39	22	47	39

Numbers in bold are more than 1.96 standard errors from the expected number (25) of rejections.

Results obtained indicate that the LOG approach had numbers of rejections larger than expected for some of the

0.70 PDIF conditions. Actually, the LOG numbers of Type I errors were generally all above the expected value in most of the small sample size conditions with the 0.70 item difficulty. However, in the 0.40 PDIF conditions most of the LOG results were above the expected nominal value regardless of the sample size. Results of logit-linear analysis indicate that a model incorporating the independent variable test length as the sole effect sufficiently explains the frequencies of acceptance and rejection ($\chi^2(14)=15.37$, $p=.353$). Slightly larger numbers of rejections were observed with the longer test conditions.

The results in Table 19 and 20 indicate that the PLR approach had good nonuniform Type I error control for all the 0.70 PDIF conditions. In the 0.40 item difficulty conditions only one condition was found to have an inflated Type I error rate. The logit-linear results indicate that a model including the interaction between PDIF and test length as well as the main effects of PDIF and test length was enough to adequately explain the variation observed in the statistical decision ($\chi^2(12)=8.66$, $p=.731$). In averaging the frequencies across the independent variables not included in the model, the shorter test length conditions were found to have slightly larger numbers of rejections in the PDIF 1 conditions, whereas for the longer test length conditions the opposite

was observed (see Table 21). However, in all cases, except one, the empirical number of Type I errors fell within expected bounds.

Table 21: PLR average number of Type I errors for the test length by PDIF condition interaction

PDIF	Test length	
	25 items	50 items
1	23.25	28.25
2	28	22.5

The numbers of rejections for both MAN(C)OVA (MCA and MCO) approaches, as expected, were often larger than would have been anticipated with a true null hypothesis (see Table 20). This is especially noticeable in the 0.70 item difficulty conditions. It is, however, possible to transform the numbers of rejections obtained with the nonuniform MAN(C)OVA into Adjusted Numbers of Rejections (ANR). In doing this, the results for the MCO and MCA are approximately comparable to those of the LOG and PLR approaches. These adjusted values may be computed by removing from the nonuniform effect, the nonuniform effect that is not expected by chance. These adjusted values are computed as follows:

$$ANR = NN_{obs} - NN_{ec} , \quad (43)$$

where

- ANR : is the adjusted number of rejections for the nonuniform test of hypothesis;
- NN_{obs} : is the observed number of significant nonuniform test of hypothesis;
- NN_{ec} : is the number of nonuniform effects that are not expected by chance.

The NN_{ec} are given by

$$NN_{ec} = NB_{obs} - NB_{exp} , \quad (44)$$

where

- NB_{obs}: is the number of times that both the uniform and nonuniform test of hypotheses are significant in 500 replications;
- NB_{exp}: is the expected number of times that both a significant uniform and nonuniform test of hypotheses occur by chance in 500 replications.

The expected probability of having both a significant uniform and nonuniform test of hypotheses by chance is given by

$$NB_{exp} = (P_{(U)obs} * P_{(M)exp}) * 500 , \quad (45)$$

where

$$P_{(U)obs} = \frac{NU_{(obs)}}{Nrep} , \quad (46)$$

and

$$P_{(N)exp} = \frac{NN_{(exp)}}{Nrep} = .05 . \quad (47)$$

Where

- $P_{(U)obs}$: The observed probability of having a significant group main effect for that condition.
- $P_{(N)exp}$: The expected probability of having a significant group by ability interaction when NO PDIF is present.
- $NU_{(obs)}$: The observed number of significant uniform tests of hypothesis in the NO PDIF conditions.
- $NN_{(exp)}$: The expected number of significant nonuniform test of hypothesis for the NO PDIF conditions.
 $NN_{(exp)} = 25$ because $25/Nrep=0.05$.
- $Nrep$: The number of replications, $Nrep=500$.

Following is an example of the computation of the ANR results using equation 43. The computational example is shown for the 0.70 PDIF 1 condition having a short test and a small sample size. For this condition, the number of observed nonuniform rejections (NN_{obs}) is 42 (see Table 19) and the number of uniform rejections (NU_{obs}) is 121 (see Table 14). Finally, the number of observed rejections for both test of hypotheses (NB_{obs}) was found to be equal to 18. This number was computed by counting the number of times that both the observed uniform and nonuniform test were significant. From these results, the ANR is calculated as follows:

$$ANR = 42 - (18 - (\frac{121}{500} * \frac{25}{500}) * 500) = 30.05$$

This value of ANR (30.05) was rounded to a value of 30 that corresponds to the value reported in Table 22.

The ANR for the MCA and MCO approaches were computed using equation 43 for the PDIF conditions. The MCA and MCO ANR results for the PDIF conditions are shown in Tables 22 and 23 for the 0.70 and 0.40 PDIF conditions, respectively. In both tables the ANR were rounded to the nearest integer.

Table 22: ANR results for the 0.70 PDIF conditions

PDIF Condition	Test Length	Sample Size	Statistics	
			MCA	MCO
1	25 items	500/250	30	34
		1500/750	31	36
	50 items	500/250	40	39
		1500/750	21	28
2	25 items	500/250	34	25
		1500/750	35	28
	50 items	500/250	29	39
		1500/750	30	30

Numbers in bold are more than 1.96 standard errors from the expected number (25) of rejections.

Table 23: ANR results for the 0.40 PDIF conditions

PDIF Condition	Test Length	Sample Size	Statistics	
			MCA	MCO
1	25 items	500/250	41	20
		1500/750	37	17
	50 items	500/250	29	25
		1500/750	37	27
2	25 items	500/250	31	24
		1500/750	31	18
	50 items	500/250	29	19
		1500/750	30	26

Numbers in bold are more than 1.96 standard errors from the expected number (25) of rejections.

Again, these adjusted scores were computed only for the PDIF conditions on the MCA and MCO nonuniform test of hypothesis. It should be stressed that the adjusted results provide only an estimate of the Type I errors of a true test of a null hypothesis and that they should be interpreted as such.

The computed results obtained indicate that some of the ANR for the test of the nonuniform hypothesis with the MCA approach were slightly above the expected values for the 0.70 and the 0.40 difficulty conditions. A logit-linear analysis on these adjusted results indicate that a model incorporating only the dependent variable (i.e. statistical decision) sufficiently explained the variation observed in the results ($\chi^2(15)=12.49$, $p=.642$).

For the MCO approach, the ANR for the tests of nonuniform hypothesis were slightly above the expected values in some of the 0.70 PDIF conditions. However, for the 0.40 PDIF conditions, the ANR for the MCO approach were all within the expected confidence interval. The results of the logit-linear analysis reveal that a model with the independent variable item difficulty sufficiently explains the variations ($\chi^2(14)=11.86$, $p=.618$). It appears that a larger number of rejections were found with the easier item conditions.

CHAPTER 6

Discussion

The results presented in the previous chapter are discussed below. First the findings pertaining to the tests of the uniform hypothesis (i.e. uniform PDIF) are outlined when no differences between the groups were introduced (i.e. NO PDIF conditions). These are followed by the results for when some uniform differences between the groups were simulated in the alternatives of the item (i.e. PDIF conditions). Next, the results for the nonuniform tests of hypothesis are discussed. For the MCA and MCO tests of nonuniform hypothesis, the adjusted numbers of rejections (ANR) for the PDIF conditions are discussed instead of their observed numbers of rejections. This was done because, as mentioned earlier, the observed values for the tests of the nonuniform hypothesis with the PDIF conditions were expected to yield larger numbers of rejections than the nominal values, and also because these results are not directly comparable to the LOG and PLR results. Because the results of the nonuniform tests of hypothesis for the NO PDIF and PDIF conditions correspond to a test of the null hypothesis, these results are discussed concurrently. Once these results are examined, the approaches are contrasted in order to ascertain which statistical method(s) seem(s) to perform well in a given

condition. Finally, a broader discussion of the results in relation to other findings is presented.

Test of uniform PDIF

NO PDIF conditions

The numbers of rejections for the statistics were, in general, within the expected nominal range (see Table 12). In some conditions, the empirical Type I error rates were slightly above the expected nominal values. However, this was not too bothersome because they were all within plus or minus three standard errors of a proportion.

The logit-linear analyses results showed that the numbers of rejections for the LOG, MCA, MCO, and PLR approaches were not affected by any of the predictors examined in this study. Indeed, the uniform tests of hypothesis for these statistics performed just as well in the detection of NO PDIF with short and long tests, small and large samples, and with easy and difficult items. It was actually expected and hoped that the independent variables would not have any effect on the statistical decision in the NO PDIF conditions.

In contrast, the logit-linear analysis results for the DS approach revealed that a test length-by-item difficulty interaction seemed to significantly influence the number of rejections. This significant interaction is difficult to explain, mainly because it may not be

necessary to conduct such an analysis on frequencies that are generally within the expected nominal range. However, it was done and a significant interaction effect was found for the DS approach. This effect needs to be further investigated before it can be adequately explained.

PDIF conditions

Four of the five statistics, namely the LOG, MCA, MCO, and PLR approaches, seemed to perform equally well for the tests of the uniform PDIF hypothesis. The numbers of rejections for these approaches were found to be very similar (see Tables 14 and 15). The results of the logit-linear analyses further confirmed this observation. The logit-linear analyses results indicate that sample size, item difficulty, and the location of PDIF all influenced the probability of rejections of the LOG, MCA, MCO, and PLR approaches. In contrast, the numbers of rejections in the DS approach were influenced by sample size, test length, and the location of PDIF. Possible explanations for these effects are provided in the following paragraphs. Each independent variable is discussed separately.

Item difficulty. In this study, PDIF was introduced by varying the probability of the alternatives. The amount by which the alternatives were varied was identical for

the two difficulty conditions. These differences which were introduced in the alternatives are shown in Table 24. These differences are expressed in terms of differences that were introduced in the focal group.

Table 24: Differences in proportions and b parameters introduced in the focal group

PDIF conditions	Option	Difference in proportion		Difference in b parameter	
		Difficulty		Difficulty	
		0.70	0.40	0.70	0.40
PDIF 1	1-2-3** 4*	+.050	+.050	+.941	+.528
		-.050	-.050	-.359	-.215
PDIF 2	1	+.017	+.017	+.296	+.181
	2	+.017	+.017	+.342	+.151
	3	+.017	+.017	+.435	+.421
	4*	-.050	-.050	-.359	-.215

* corresponds to the correct alternative

** 1-2-3 indicates that only one of these distractors was changed

For example, the correct alternative for the PDIF 1 or PDIF 2 conditions was made harder for the focal group by reducing the probability of selecting the correct answer by 0.05. However, in Table 24 it is observed that although equal differences in proportions were introduced in the alternatives of the item, this translated into introducing larger differences in the b parameters for the easier item. That is, in the PDIF 1 conditions introducing a difference in proportion of 0.05 in the correct alternative translated into a difference in b parameters of .215 for the difficult item ($p=0.40$) and of .359 for

the easy item ($p=0.70$). This occurs because the b parameter is a function of both the z scores corresponding to the probability of choosing alternative e to item i and the biserial correlation value of the alternative (see equation 42). Because the b parameters were used to generate the data, the greater power observed with the easier item for the LOG, MCA, MCO, and PLR approaches is simply the result of a greater effect size. It is, however, important to note that the DS approach was not able to detect that difference and that the DS approach appears to be less powerful than the other approaches.

Location of PDIF. The power of the statistics was found to be significantly affected by how PDIF was introduced in the item (i.e. location of PDIF). Of the two conditions investigated, PDIF 1 was easier to detect than PDIF 2. This effect is easily understood by comparing the manner in which PDIF was introduced in each condition. In PDIF 1, the proportion of only one distractor was made more attractive to the focal group while in PDIF 2 all the distractors' proportions were made more attractive to the focal group. The differences in proportions of the distractors in the PDIF 2 conditions was equivalent to one third of the differences in the distractors in PDIF 1. In light of this, it is not surprising to get greater power with the PDIF 1 conditions because the effect sizes in the

distractors are larger than in the PDIF 2 conditions. In conclusion, the differences in effect size among the distractors accounts for the differences in power observed for all the statistics.

Sample size. A significant sample size effect was found for all the approaches. The statistics were found to be more powerful with the larger sample size conditions. This is not surprising because it is well known that the statistical power is directly related to sample size, with greater power being associated with a larger sample size (Stevens, 1986; Tatsuoaka & Lohnes, 1988). In summary, the sample sizes investigated in this study yielded greater power of rejections for the larger sample size conditions.

Test length. Finally, the DS approach was the only method found to be significantly affected by the length of the test. Actually, true differences should be easier to detect with longer tests because they should yield more reliable scores. This effect was observed with the DS approach. This probably occurs because the item weights are expected to be more stable with longer tests as compared to shorter tests and, therefore, any differences that do exist should be detected more often.

It is, however, important to mention that the other statistics were not affected by a change in test length.

Test length is known to affect the accuracy of the ability estimates (Swaminathan & Rogers, 1990), however the ranges in test length investigated in this study were not found to significantly affect the LOG, MCA, MCO, and PLR approaches in the detection of uniform PDIF. One plausible explanation for this might be that the range of test lengths investigated was not large enough to significantly affect the accuracy of the ability estimates for the LOG, MCA, MCO, and PLR approaches. This hypothesis deserves to be further investigated. As well, the two-parameter model that was used to generate the data does not yield data having much measurement error. Perhaps the test length would have made a difference if the data had been generated using a three-parameter logistic model.

Test of nonuniform PDIF

The results of the nonuniform tests of hypothesis are discussed in this section. The results for the NO PDIF and PDIF conditions are discussed concurrently because these tests correspond to a test of the null hypothesis, except for the observed MCA and MCO tests of nonuniform hypothesis with the PDIF conditions. However, only the MCA and MCO ANR results for the PDIF conditions are discussed here. The observed PDIF results for the MCA and MCO tests of nonuniform hypothesis were not of interest because they did not correspond to a true test of the null hypothesis.

For clarity each approach is discussed under it's own heading.

LOG approach

The LOG approach was found to have poor Type I error rates for the NO PDIF and PDIF conditions. Logit-linear analysis results indicated that sample size appeared to affect the number of significant nonuniform tests of hypothesis for the NO PDIF conditions, and that test length affected the number of significant nonuniform tests of hypothesis for the PDIF conditions. It is surprising that different models were observed to be influential for the PDIF and the NO PDIF conditions. This was especially surprising, because for these conditions the test of nonuniform hypothesis corresponded to a test of the null hypothesis. This discrepancy suggests that one or both effects may have occurred by chance.

The effect of sample size observed with the NO PDIF conditions indicated that the LOG approach in the smaller sample size conditions tended to misclassify an item as having PDIF when in fact NO PDIF was present. This may have resulted because it is more difficult to replicate equivalent response distributions across the alternatives between the focal and reference groups with smaller sample sizes. Typically, the larger the size of the sample the better the chances are of having proportions that resemble

those shown in Table 25. More specifically, the probability of obtaining proportion values identical to those found in Table 25 is much larger for large sample size conditions than for small sample size conditions. More differences may occur between the groups with smaller sample size conditions because it is harder to replicate the expected frequencies.

Table 25: Example of expected proportions for an item

Alternatives			
1	2	3	4
.15	.10	.05	.70

Test length is known to have an effect on the reliability of the total test scores. More reliable test scores are usually found with longer tests (Crocker & Algina, 1986; Swaminathan & Rogers, 1990). Therefore, it is expected that larger numbers of rejections would occur with shorter tests. However, for the PDIF conditions, significantly larger numbers of rejections were observed with the longer test conditions. This observation is contradictory to usual observations. One reason for this might be due to the fact that the PDIF introduced was not perfectly uniform as intended. Indeed, Figures 4, 5, 6, 10, 11, and 12 indicate that a very slight nonuniform PDIF was also introduced in the alternatives of the item. These nonuniform differences, even though they were very small,

would be easier to detect with longer a test as observed with the data.

However, because the test of nonuniform hypothesis for PDIF and NO PDIF conditions yielded different significant effects, it is still uncertain if these effects would be replicated with other data sets. Therefore, further research should be conducted before any conclusions might be reached about these effects.

MCA approach

In general, the frequencies of rejections for the nonuniform tests were within the expected nominal range for the MCA approach for the NO PDIF conditions. Whereas, some of the nonuniform tests of hypothesis were above the expected nominal for the PDIF conditions. This observation is made from the observed results of the NO PDIF conditions and the adjusted results of the PDIF conditions. The logit-linear analysis results show that the adjusted MCA results of the nonuniform tests of hypothesis for the PDIF conditions were not found to be affected by any of the independent variables. However, the logit-linear analysis results for the NO PDIF indicated that the observed frequencies of rejections were affected by a sample size-by-test length interaction. These results are again difficult to interpret since it is probably not necessary to conduct a logit-linear analysis on values

which are mainly within the expected nominal range. However, this interaction cannot be ignored and it would be best to wait for further research before any definite conclusions might be reached.

MCO Approach

The observed frequencies of rejections for the NO PDIF conditions were above the nominal range for the 0.70 item difficulty conditions, while for the 0.40 item difficulty conditions they tended to be within the expected nominal range. Similar results were also found for the ANR of the PDIF conditions. Therefore, the numbers of rejections for the nonuniform tests of hypotheses were larger than expected in the easier item conditions. This may have resulted because more variability between the ability of the subjects and their performance on the item probably occurs with easier items.

To understand how more variability can occur with the easier items, let us first compare the biserial correlation values of these two studied items. The biserial correlation for the harder item was 0.61 as compared to 0.40 for the easier item. The easier item discriminates less well than the harder item. This indicates that there is less of a relationship between the ability of the subjects and their performance on the item with the easier item. This may in turn have increased the

numbers of significant differences that occurred by chance.

PLR approach

In general it can be said that the numbers of rejections using PLR were all within sampling error for the nonuniform tests of hypothesis in both conditions (i.e. PDIF and NO PDIF). This was somewhat expected because no nonuniform differences had been introduced in the alternatives of the item. The logit-linear analysis results indicated that a model including a test length-by-item difficulty interaction, and a model including a test length by location of PDIF interaction explained the frequencies of rejections in the NO PDIF and PDIF conditions, respectively. These results are surprising and are difficult to interpret, because the logit-analyses were computed on results that were clearly within the expected nominal range. Because the frequencies of rejections were clearly within the expected nominal range, it may not be necessary to conduct a logit-linear analysis on such results. However, the analyses were done and for the PDIF and NO PDIF conditions significant interactions were found to explain the variations in the results. It is worthwhile to mention that different models were found to explain the variations observed for the NO PDIF and the PDIF conditions. This observation alone suggests that

these significant interactions might not be found with other data sets. However, before any definitive conclusions may be reached more research needs to be conducted.

Contrasting the approaches

The results of this investigation showed that the LOG, MCA, MCO, and PLR results were similar for the uniform tests of hypothesis in the PDIF and NO PDIF conditions. Actually, all these methods, namely the LOG, MCA, and PLR have been shown to be related to one another in two-way contingency tables. Goodman (1979, 1981), Knapp (1978), van der Heijden, de Falguerolles, and de Leeuw (1989), and van der Heijden and de Leeuw (1985) have shown that in two-way contingency tables the log-linear (LOG) method yields estimates that are approximately the same as the weights derived from canonical correlation analysis. Note that the canonical correlation analysis is similar to the MCA approach employed in this study. In addition, it was demonstrated by Fienberg and Meyer (1983) and Swafford (1980) that a logistic regression model can be represented by an equivalent log-linear model when all the variables are categorical. Therefore, there is some evidence that the log-linear, logistic regression, and the canonical correlation (i.e. corresponding to MCA in this study) all yield approximately equivalent results in two-way

contingency tables. Finally, the MCO approach has the same relationship with the MCA approach that the PLR has with LOG approach. That is, the only difference between these two approaches is that the measure of ability with the MCO approach is continuous instead of being categorical as in the MCA approach.

The results of this study indicate that these relationships are evident, but only for the uniform tests of hypothesis. Indeed, the numbers of rejections were similar with the uniform tests of hypothesis for the PDIF and NO PDIF conditions, while the nonuniform tests of hypothesis results were somewhat different among the approaches investigated. Actually, the PLR Type I error rates were the only one which tended to be within the expected nominal range for the PDIF and NO PDIF conditions. The Type I error rates for the nonuniform tests of hypothesis of the LOG approach, however, tended to be more affected by some of the changes in the independent variables. Actually, the LOG results were affected by the length of the test in the PDIF conditions and by the size of the sample in the NO PDIF conditions. It should be noted that the results of this study do not permit a clear indication as to which of these effects, namely test length and sample size, influence the Type I error rates of the LOG approach. More specifically, the effects of test length and sample size are not clear

because in the PDIF and NO PDIF conditions different effects were found to affect the nonuniform Type I error rates. The Type I error rates for the MCO approach were found to be affected by the length of the test for the nonuniform NO PDIF and PDIF conditions. Finally, the MCA results indicated that the Type I error rates were within the expected confidence intervals with the NO PDIF conditions, but that some of the nonuniform adjusted MCA Type I error rates for the PDIF conditions were slightly larger than expected. It should be noted that because the nonuniform MCA and MCO Type I error rates for the PDIF conditions were approximated values, the conclusions regarding the effectiveness of these two approaches were based mainly on the NO PDIF conditions. Under the NO PDIF conditions the MCA seemed to maintain a better control on the Type I error rates than the MCO approach, at least in the conditions investigated in this study.

It is somewhat surprising to have found that the approaches that tended to have good Type I error rates were the MCA and PLR approaches. (Note that this statement does not include the results of nonuniform tests of hypothesis for the MCA approach). This was unforeseen mainly because the PLR uses a continuous measure of ability while the measure of ability for the MCA is categorical. Actually, the number of Type I errors for the PLR approach were expected to be more comparable with

those of the MCO approach because they both use a continuous measure of ability.

The DS approach did not compare favourably to the other approaches. The power of the DS approach was much lower than all the other approaches, but it did have good Type I error control. These results, at first glance, cast serious doubt on the usefulness of this approach. However, after investigating possible reasons for obtaining lower power than expected, it is believed that the inclusion of only one item exhibiting PDIF in the total test may explain the low power observed. To understand how this may be, one has to fully understand that the purpose of the DS approach is to examine PDIF within a multidimensional framework. This is similar to the conceptual framework under which Ackerman (1992), and Shealy and Stout (1989) study DIF. For simplicity, the discussion is limited to a two-dimensional perspective, where one dimension represents the intended measure of ability and the second represents what Ackerman (1992) refers to as the nuisance ability. The nuisance ability is not intended to be measured by the test and if the groups differ on this dimension (i.e. nuisance ability) the potential for bias exists. It was thought that the weights of the second dimension derived by Dual Scaling for the focal and reference groups could be compared to assess PDIF. However, the inclusion of only one item in the test as

having PDIF may not have created a significant second dimension, especially given that the amount of PDIF was quite small. If more items with PDIF were included in the test it is possible that adequate power may have resulted with the DS approach. This certainly deserves to be further investigated.

It is noteworthy to indicate that the nonuniform tests of hypothesis for the MCA approach were not affected by the difficulty of the item as was the MCO approach. The only difference between these two approaches was the scale of the ability measure. For the MCA approach the measure of ability is categorical (i.e. ability groups based on the total test scores) while for the MCO approach it is continuous (i.e. total test score). It was earlier suggested that the numbers of significant nonuniform tests of hypothesis for the MCO approach were larger for the easier item because more variability between the ability of the subjects and their performance on the item might exist with the easier item. It, however, seems worthwhile to categorize the measure of ability with the easier item as it probably tends to lessen that variability. In doing this, the numbers of rejections seemed to reduce to acceptable values. However, it is still not yet clear as to why the MCA was affected by this effect but the PLR approach was not.

Comparing the results with other studies

The results of this investigation are not directly comparable to other IRD studies because similar issues have not been addressed. However, some points deserve to be mentioned in this section.

The logistic regression procedure has been shown by Swaminathan and Rogers (1990) to have potential in detecting uniform and nonuniform DIF. This study shows that the PLR can also easily be used in the assessment of uniform and nonuniform PDIF. The PLR was found to have good Type I error control as well as adequately detecting uniform PDIF. However, it should be remembered that the performance of the PLR was assessed based on item responses generated with a PLR model.

Mazor, Kanjee, and Clauser (1993) have shown that, in the dichotomous case, the logistic regression approach can be employed with multiple ability estimates. It is believed that the same approach can be extended to the polytomous case. However, the use and efficiency of the PLR approach with multiple ability estimates has yet to be investigated.

Finally, the PLR approach investigated here is not directly comparable to Wilson, Spray, and Miller's (1993) logistic regression approach because they basically analyzed polytomous response data in a dichotomous manner. As well their approach requires that the item responses be

ordinal. However, the similarity of the procedures should not be overlooked and their results pertaining to the detection of nonuniform PDIF should be discussed as they may lend some support for the use of the PLR approach.

Wilson, Spray, and Miller (1993) investigated the usefulness of the logistic regression procedure in its assessment of nonuniform PDIF. To do this, they introduced nonuniform PDIF by varying the a (discrimination) and b (difficulty) parameters of the graded response model (Muraki, 1991) which was used to generate their data. They considered four conditions of nonuniform PDIF. For three conditions only the a parameter was altered and for the other condition only the b parameter was varied. They labelled all these conditions as nonuniform PDIF conditions. They then used the logistic regression approach in order to detect these four types of nonuniform PDIF. Obviously, they found that the logistic regression procedures they used flagged the fourth condition (i.e. where only b was varied) as uniform PDIF and based on this observation they concluded that the logistic regression procedure incorrectly detects nonuniform PDIF as uniform PDIF. This is only true if nonuniform PDIF is defined by a group-by-alternative interaction. Otherwise, if nonuniform PDIF is only defined in terms of a group-by-ability interaction the approach, as is traditionally the case,

used by Wilson, Spray, and Miller (1993) correctly identifies uniform and nonuniform PDIF.

In their study, they also investigated the effect of sample size on the power of detection. They found that the larger sample size (i.e. 2000 examinees in both groups) conditions were more powerful than the smaller one (i.e. 500 examinees in both groups). However, they noted that the differences they observed were not large. The power results for the PLR approach in this study were similar in that sample size was also found to have an effect on the detection rates.

The MCA approach which showed potential by Gessaroli (1985) has also been found in this study to work well in the detection of uniform PDIF. As well, the technique was found to have good Type I error control. In Gessaroli (1985) the MCA approach was investigated under conditions of equal sample size and underlying normally distributed abilities. In this investigation the underlying subjects' abilities response were also assumed to be normally distributed but unequal sample size conditions were examined. Under these new conditions the MCA approach was found to perform well in the detection of uniform PDIF.

CHAPTER 7

Conclusions

The conclusions pertaining to this investigations are presented in this section. This chapter is divided into three sections. The first section is concerned with the conclusions of the uniform test of hypothesis, the second section addresses the conclusions pertaining to the nonuniform test of hypothesis, and finally the last section present some overall conclusions. It should, however, be remembered that no attempt is made to generalize beyond the scope of this study. After the conclusions, some recommendations for future research are also discussed.

Uniform tests of hypothesis

All the approaches, except for the DS approach, were found to adequately detect uniform PDIF. The DS approach was found less powerful than the other approaches. The power of detection for the LOG, MCA, MCO, and PLR approaches were all found to be affected by the difficulty of the item, the size of the samples, and the location of PDIF. The DS approach was found to be also affected by the size of the samples and the location of PDIF as well as the length of the test.

All the approaches were found to have good Type I error rates when NO PDIF was introduced in the

alternatives of the item. The numbers of rejections were mainly within the expected nominal range.

Nonuniform tests of hypothesis

The numbers of Type I errors for the NO PDIF conditions were mainly within the expected nominal range for the MCA and PLR, while the other two approaches (LOG and MCO) had larger than expected numbers of Type I errors. For the PDIF conditions, the PLR Type I error results were all found to be within the expected nominal range, while the LOG and adjusted MCA and MCO had some inflated Type I error rates. The PLR method was the only one which was found to maintain adequate Type I error control in both the NO PDIF and PDIF conditions. Because the MCA and MCO Type I error rates were approximated for the nonuniform PDIF conditions, conclusions for these approaches are based only on the performance of these approaches in the NO PDIF conditions. Based on this the MCA was the only other approach found to maintain adequate Type I error rates in the nonuniform conditions.

For most of these approaches, the results of the logit-linear analyses, however, suggested that the dependent variable (i.e. statistical decision) was not the sole predictor of the variation found in the frequencies of rejections. These results are, however, not conclusive because they were computed on rejection numbers that were

clearly within the expected nominal range and, more importantly, the effects observed in the NO PDIF conditions were not replicated in the PDIF conditions. Actually, the only Type I error rates that were clearly found to be related to one of the independent variables were obtained with the MCO approach. The MCO Type I error rates for the PDIF and NO PDIF conditions were found to be affected by the difficulty of the item, where the easier item conditions tended to have elevated Type I rates. This was suggested to be related to the fact that this item discriminated less than the more difficult one. Therefore, based on these preliminary results it seems that the MCO Type I rates for the nonuniform test of hypothesis might be affected by the quality of the item (i.e. discrimination).

Overall conclusions

In summary, it can be concluded that the MCA and PLR were the only approaches that were found to have potential in the analysis of PDIF. When no differences in the alternatives of the item existed, the MCA and PLR approaches had good Type I error control for the tests of uniform and nonuniform PDIF. Similarly, the MCA and PLR approaches were found to adequately detect uniform PDIF when it had been introduced in the alternatives of the item. The PLR was also found to maintain a good control on

the number of Type I errors for the tests of nonuniform hypotheses. The observed MCA results were, however, found to have elevated Type I errors for the nonuniform test of hypothesis when uniform PDIF had been introduced in the alternatives of the item. Although, the ANR results for the MCA approach improved the Type I error rates some of the values remained elevated for the nonuniform tests of hypotheses when uniform PDIF had been introduced. Because the ANR results give an approximation of the true test of the null hypothesis, they cannot be used to assess the real effectiveness of this approach. Therefore, no conclusions were drawn from these ANR results. To better investigate the Type I error control of the MAN(C)OVA approaches a non-logistic model could perhaps be employed to simulate the data.

If the purpose of the PDIF study is to only identify uniform PDIF, then the LOG and MCO approaches may also be used because they also maintained good Type I error control as well as adequately detecting uniform PDIF.

Finally, the DS approach was found to be less powerful than the other approaches in the detection of uniform PDIF although it was found to maintain good Type I error control when no PDIF was introduced. This lack of power is thought to be related to the low number of items exhibiting PDIF in the test (i.e. 1 item). It is believed that the adequacy of this technique should be investigated

when more items in the test exhibit PDIF. If this technique proves to be useful in these conditions it would have the advantage of studying PDIF from a multidimensional perspective. Also, in a sense it can be said that it would reconcile differential test performance with IRD. That is if no overall test differences are found between the groups of interest then there is no need to look for IRD.

Suggestions for future research

In light of the findings of this study, the following suggestions are offered regarding future investigations that might be undertaken in this area of research. It should be mentioned that it is implied, unless specified otherwise, that the following recommendations be made with regards to the approaches that were found in this study to have potential in the analysis of PDIF, namely, the MCA and PLR.

First, it would seem to be important to assess the effectiveness of these approaches under different conditions of test length, sample size, item difficulty, and the location of PDIF. The purpose would be to develop guidelines under which these approaches might be used.

Second, a different simulation procedure should be employed in order to better assess the accuracy of the PLR approach. In this investigation the use of a two-parameter

polytomous logistic regression model tends, at least theoretically, to favour the PLR approach results.

Third, the accuracy of the approaches should be investigated when more items exhibiting PDIF are included in the test. This is expected to affect the accuracy of the ability estimates and, therefore, the effect of test length on the accuracy of the approaches might be more important. Actually, the DS approach should especially be examined when more items in the test have PDIF because it is expected that this technique should be more powerful than was observed in this study.

Fourth, the latent variable underlying ability was assumed to be normally distributed for both the focal and reference groups. It is fairly common to assume such a distribution but in reality this may not be true. Therefore, it would be important to study the effectiveness of these approaches when the ability is not normally distributed. Also, it would be important to assess the accuracy of the techniques when the groups have unequal underlying abilities.

Fifth, in this study only uniform logit differences were introduced in the alternatives of the item. It would seem important to assess the performances of the approaches in the detection of nonuniform PDIF.

Finally, some authors (Begg and Gray, 1984; Wilson, Spray, and Miller, 1993) have cautioned that the PLR or

logistic regression approach might be problematic when used with sparse data. In these conditions the power of the statistic would be affected (Wilson, Spray, & Miller, 1993). This might cause similar problems with the LOG approach and the MCA approach. However, at this stage, it is not known how having empty cells or very small categories affects the accuracy of the approaches. This problem should be addressed in future research.

References

- Ackerman, T.A. (1991a, April). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Ackerman, T.A. (1991b). Measurement direction in a multidimensional latent space and the role it plays in bias detection. Paper presented at the 1991 International Symposium on Modern Theories in Measurement: Problems and issues. Montebello, Quebec.
- Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29(1), 67-91.
- Ackerman, T.A., & Evans, J.A. (1993, April). A didactic example of the influence of conditioning on the complete latent ability space when performing DIF analysis. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA.
- Aldrich, J.H., & Nelson, F.D. (1984). Linear Probability, Logit, and Probit Models. Beverly Hills, CA: Sage Publications.
- Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In Holland, P.W., & Wainer, H. (Eds.), Differential Item Functioning (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Baker, F.B. (1981). Log-linear, logit-linear models: A didactic. Journal of Educational Statistics, 6(1), 75-102.
- Baker, F.B., & Subkoviak, M.J. (1981). Analysis of test results via log-linear models. Applied Psychological Measurement, 5(4), 503-515.
- Bartlett, M.S. (1939). A note on tests of significance in multivariate analysis. Proceedings of the Cambridge Philosophical Society, 35, 180-185.

- Begg, C.B., & Gray, R. (1984). Calculation of polytomous logistic regression parameters using individualized regression. Biometrika, 7(1), 11-18.
- Bennet, R.E., Rock, D.A., & Kaplan, B.A. (1987). SAT differential item performance for nine handicapped groups. Journal of Educational Measurement, 24, 41-55.
- BMDP (1990). Statistical Software Manual. Volume 2. University of California press, Berkeley
- Bock, R.D. (1972). Estimating item parameters and latent ability and validity of scoring aptitude and achievement tests with weights for every choice. Psychometrika, 37, 29-51.
- Bock, R.D. (1975). Multivariate Statistical Methods in Behavioral Research. New-York, NY: McGraw-Hill.
- Burrill, L.E. (1982). Comparative studies of item bias methods. In Berk, R.A. (Ed.) Handbook of Methods for Detecting Test Bias. Baltimore, NJ: The John Hopkins University Press.
- Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. Applied Psychological Measurement, 16(2), 129-147.
- Cole, N.S. (1973). Bias in selection. Journal of Educational Measurement, 10(4), 237-255.
- Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New-York: Holt, Rinehart and Winston.
- Darlington, R.B. (1971). Another look at "cultural fairness". Journal of Educational Measurement, 8(2), 71-82.
- Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In Holland, P.W., & Wainer, H. (Eds.), Differential Item Functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Dorans, N.J., & Schmitt, A.P., & Bleistein, C.A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement, 29(4), 309-39.

- Drasgow, F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.
- Fienberg, S.E. (1980). The Analysis of Cross-Classified Categorical Data. Cambridge, MA: The MIT Press.
- Fienberg, S.E., & Meyer, M.M. (1983). Log-linear models and categorical data analysis with psychometric and econometric applications. Journal of Econometrics, 22, 191-214.
- Fisher, R.A. (1940). The precision of Discriminant Functions. Annals of Eugenics, (10), 422-429.
- Gessaroli, M.E. (1985). A Monte Carlo Investigation of the Type I Error Rates of Three Multivariate Tests Applied to Categorical Data. Unpublished doctoral dissertation, University of Toronto, Toronto.
- Gigula, Z. (1986). Grouping and association in contingency tables; An exploratory canonical correlation approach. Journal of the American Statistical Association, 81 (395), 773-779.
- Gigula, Z., & Haberman, S.J. (1988). The analysis of Multivariate contingency tables by restricted canonical and restricted association models. Journal of the American Statistical Association, 83(403), 760-771.
- Goodman, L.A. (1979). Simple models for the analysis of association in cross-classification having ordered categories. Journal of the American Statistical Association, 74 (367), 537-552.
- Goodman, L.A. (1981). Association Models and Canonical Correlation in the analysis of cross-classifications having ordered categories. Journal of the American Statistical Association, 76, 320-334.
- Green, B.F. (1991). Differential item functioning: Techniques, findings and prospects. Paper presented at the 1991 International Symposium on Modern Theories in Measurement Problems and Issues. Montebello, Quebec.

- Green, B.F., Crone, C.R., & Folk, V.G. (1989). A method for studying differential distractor functioning. Journal of Educational Measurement, 26(2), 147-160.
- Greenacre, M.J. (1984). Theory and Applications of Correspondence Analysis. London: Academic Press Inc.
- Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In the Committee on Social Adjustment (Ed.). The Prediction of Personal Adjustment, (pp. 319-348). New-York, NY: Social Science Research Council.
- Guttman, L. (1966). The principal components of scale analysis. In S.A. Stouffer, L. Guttman, E.A., P.F. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds.). Measurement and Prediction. New-York, NY: Science Editions, John Wiley and Sons Inc.
- Haberman, S.J. (1978). Analysis of Qualitative Data. New-York, NY: Academic Press.
- Healy, M.J.R., & Goldstein, H. (1976). An approach to the scaling of categorized attributed. Biometrika, 63, 219-29.
- Hills, J.R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.
- Hirschfeld, H.O. (1935). A connection between correlation and contingency. Cambridge Philosophical Society Proceedings, 31, 520-524.
- Holland, T.R., and Levi, M. (1980). Canonical correlation in the analysis of a contingency table. Psychological Bulletin, 87 (2), 334-336.
- Holland, P.W., & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. (Program Statistics Research, Technical Report No. 86-82). Princeton, NJ: Educational Testing Services.
- Horst, P. (1935). Measuring complex attitudes. Journal of Social Psychology, 6, 369-374.
- Hosmer, D.W., & Lemeshow, S. (1989). Applied Logistic Regression (pp. 216-245). New-York, NY: John Wiley and Sons.

- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417-41, 498-520.
- Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28, 321-377.
- Ironson, G.H. (1982). Use of chi-square methods and latent trait approaches for determining item bias. In Berk, R.A. (Ed.), Handbook of methods for detecting test bias (pp.117-155). Baltimore: The Johns Hopkins University Press.
- Jacobs, P.I., & Vandeventer, M. (1970). Information in wrong responses. Psychological Reports, 26, 311-315.
- Junker, B.W., & Stout, W.F. (1991). Robustness of ability estimation when multiple traits are present with one trait dominant. Paper presented at the 1991 International Symposium on Modern Theories in Measurement: Problems and Issues, Montebello, Quebec.
- Kelderman, H., & Macready, G.B. (1990). The use of log-linear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27(4), 307-327.
- Kendall, M.G., & Stuart, A. (1967). The Advanced Theory of Statistics, Volume II. London Griffin.
- Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance-testing system. Psychological Bulletin, 82 (2), 410-416.
- Kshirsagar, A.M. (1972). Multivariate analysis. New York: Dekker.
- Lancaster, H.O. (1958). The structure of bivariate distributions. Annals of Mathematical Statistics, 29, 719-736.
- Levine, M.V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.
- Linn, R.L. (1971). Fair test use in selection. Review of Educational Research, 43(2), 139-161.

- Linn, R.L. (1984). Selection bias: Multiple meanings. Journal of Educational Measurement, 21(1), 33-47.
- Linn, R.L., & Drasgow, F.K. (1987). Implications of the golden rule settlement of test construction. Educational Measurement: Issues and Practice, 6, 13-17.
- Linn, R.L., Levine, M.V., Hasting, C.N., & Wardrop, J.L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5(2), 159-173.
- Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Mantel, N. (1963). Chi-square tests with one degree of freedom extensions of the Mantel-Haenszel procedure. American Statistical Association Journal. 690-700.
- Master, G.N. (1984). Constructing an item bank using partial credit scoring. Journal of Educational Measurement, 21(1), 19-32.
- Maung, K. (1941). Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of scottish school children. Annals of Eugenics, 11, 189-223.
- Mazor, K.M., Kanjee, A., & Clauser, B.E. (1993, April). Using logistic regression with multiple ability estimates to detect differential item functioning. Paper presented at the annual meeting of National Council on Measurement in Education, Atlanta, GA.
- McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale New-Jersey: Lawrence Erlbaum Associates, Publishers.
- McDonald, R.P., Torii, Y., & Nishisato, S. (1979). Some results on proper eigenvalues and eigenvectors with applications to scaling. Psychometrika, 44, 211-227.
- Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7(2), 105-118.
- Mellenbergh, G.J. (1988, April). Experimental research on the explanation of item bias. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Muraki, E. (1991). A generalized partial credit model: Application of an EM algorithm. Unpublished paper.
- Nishisato, S. (1980). Analysis of Categorical Data: Dual Scaling and its Application. Toronto, Canada: University Press of Toronto.
- Nishisato, S., & Lawrence, D.R. (1989). Dual scaling of multiway data matrices: Several variants. In R. Coppi and S. Bolasco (Eds.), Multiway Data Analysis (pp. 317-326). Amsterdam, North Holland: Elsevier Science Publishers.
- O'Neill, M.E. (1978a). Asymptotic distribution of the canonical correlation from the contingency tables. Australian Journal of Statistics, 20(1), 75-82.
- O'Neill, M.E. (1978b). Distribution expansion for canonical correlation from contingency tables. Journal of the Royal Statistical Society B, 40, 303-312.
- O'Neill, M.E. (1981) A note on the canonical correlations from contingency tables. Australian Journal of Statistics, 23(1), 58-66.
- Oshima, T.C., & Miller, M.D. (1991, April). Multidimensionality and item bias in item response theory. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Pillai, K.C.S. (1955). Some new test criteria in multivariate analysis. Annals of Mathematical Statistics, 26, 117-121.
- Richardson, M., & Kuder, G.F. (1933). Making a rating scale that measures. Personnel Journal, 12, 36-40
- Samejima, F. (1979). A new family of models for the multiple-choice item. (Office of Naval Research Reports No. 79-4). Knoxville, TE: University of Tennessee.
- SAS Institute INC. (1990) SAS User's Guide: Statistics, version 6.07. Cary, NC: SAS Institute Inc.
- Shealy, R., & Stout, W. (1989, April). A Procedure to Detect test bias present simultaneously in several items. (Technical Report 91-3-ONR). Champaign IL University of Illinois.

- Shealy, R., & Stout, W. (1991). A procedure to detect test bias present simultaneously in several items. Paper presented at the annual meeting of the American Educational Research Association: San Francisco, CA.
- Shealy, R.T., & Stout, W.E. (1993). An item response theory model for test bias and differential test functioning. In Holland, P.W., & Wainer, H. (Eds.). Differential Item Functioning, (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Shepard, L.A. (1982). Definition of bias. In Berk, R.A. (Ed) Handbook of Methods for Detecting Test Bias. Baltimore, MD: John Hopkins University Press.
- Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. Journal of Educational Statistics, 9(2), 93-128.
- Shepard, L., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9(2), 93-128.
- Siegel, I.E. (1963). How intelligence tests limit understanding of intelligence. Merrill-Palmer Quarterly of Behaviour and Development, 9(1), 39-56.
- SPSS (1990). SPSS User's Guide, 4.1 Edition. Chicago, IL: SPSS Inc.
- Srikantan, K.S. (1970). Canonical association between nominal measurements. Journal of the American Statistical Association, 65, 284-292.
- Stevens, J. (1986). Applied Multivariate Statistics for the Social Sciences. Hillsdale, New-Jersey: Lawrence Erlbaum Associates, Publishers.
- Swafford, M. (1980). Three parametric techniques for contingency table analysis: A nontechnical commentary. American Sociology Review, 45, 664-690.
- Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

- Sympson, J.B. (1983). A new IRT model for calibrating multiple choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.
- Tabachnick, B.G., Fidell, L.S. (1989). Using Multivariate Statistics, 2nd Edition. New York, NY; Harper and Row, Publishers.
- Tatsuoka, M.M., & Lohnes, P.R. (1988). Multivariate Analysis: Techniques for Educational and Psychological Research, 2nd Edition. New-York, NY: Macmillan Publishing Company.
- Tenenhaus, M., & Young, F.W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. Psychometrika, 50(1), 91-119.
- Thissen, D.M. (1976). Information in wrong responses to the raven progressive matrices. Journal of Educational Measurement, 13(3), 201-214.
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49(4), 501-519.
- Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement, 26(2), 161-176.
- 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 (pp. 147-169). Hillsdale, NY: Lawrence Erlbaum Associates, Inc.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland, P.W., & Wainer, H. (Eds.). Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Thorndike, R.L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8(2), 63-70.
- van der Heijden, P.G.M., de Falguerolles, A., & de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear. Applied Statistics, 38(2), 249-292.

- van der Heijden, P.G.M., & de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. Psychometrika, 50(4), 429-447.
- Veale, J.R., & Foreman, D.I. (1983). Assessing cultural bias using foil response data: cultural variation. Journal of Educational Measurement, 20(3), 249-258.
- Wainer, H. (1989). The future of item analysis. Journal of Educational Measurement, 26(2), 191-208.
- Welch, C.J., Ackerman, T.A., & Doolittle, A.E. (1987, April). An examination of statistical procedures for detecting cross-cultural differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.
- Welch, C., & Hoover, H.D. (1991, April). Procedures for extending item bias techniques to polytomously scored items. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Wilson, A.W., Spray, J.A., & Miller, T.R. (1993, April). Logistic regression and its use in detecting nonuniform differential item functioning in polytomous items. Paper presented at annual meeting of the National Council on Measurement in Education, Atlanta, GA.
- Wright, D.J. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Zwick, R., & Cramer, E.M. (1986). A multivariate perspective on the analysis of categorical data. Applied Psychological Measurement, 10 (2), 141-145.

Appendix A

* ITEM 1 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1207	-0.0899	-0.1458
2	0.1362	-0.2398	-0.3765
3	0.1238	-0.1657	-0.2670
* 4	0.6192	0.2390	0.3046
OTHER	0.0000		

VALID CASES: 323
CASES WITH MISSING VALUES: 0

* ITEM 2 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0557	-0.2252	-0.4595
* 2	0.5882	0.3420	0.4325
3	0.1207	-0.2168	-0.3516
4	0.2353	-0.1214	-0.1674
OTHER	0.0000		

VALID CASES: 323
CASES WITH MISSING VALUES: 0

* ITEM 3 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
* 1	0.4118	0.4809	0.6082
2	0.2786	-0.2031	-0.2711
3	0.1207	-0.3237	-0.5250
4	0.1889	-0.1027	-0.1486
OTHER	0.0000		

VALID CASES: 323
CASES WITH MISSING VALUES: 0

 * ITEM 4 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.3622	-0.1549	-0.1986
2	0.2167	-0.0474	-0.0665
* 3	0.3189	0.2163	0.2823
4	0.1022	-0.1567	-0.2663
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 5 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
* 1	0.5635	0.4288	0.5399
2	0.0433	-0.1579	-0.3497
3	0.1765	-0.2203	-0.3240
4	0.2167	-0.2343	-0.3289
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 6 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.2910	-0.1671	-0.2214
2	0.1517	-0.2626	-0.4010
* 3	0.4830	0.3171	0.3975
4	0.0743	-0.1214	-0.2265
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 7 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1424	-0.2669	-0.4142
* 2	0.6904	0.3429	0.4496
3	0.1084	-0.1807	-0.3019
4	0.0588	-0.1213	-0.2434
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 8. *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1703	-0.1974	-0.2929
* 2	0.4396	0.2518	0.3169
3	0.2446	-0.2009	-0.2750
4	0.1455	0.0582	0.0898
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 9 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0805	-0.3048	-0.5553
2	0.0960	-0.2237	-0.3869
3	0.1672	-0.2447	-0.3647
* 4	0.6563	0.4182	0.5398
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 10 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1796	-0.0613	-0.0897
2	0.2724	-0.2838	-0.3804
* 3	0.4830	0.2911	0.3649
4	0.0650	-0.1679	-0.3266
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 11 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0650	-0.2305	-0.4482
2	0.0991	-0.1916	-0.3285
* 3	0.5666	0.2924	0.3683
4	0.2693	-0.1088	-0.1461
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 12 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0372	-0.0649	-0.1513
* 2	0.3282	-0.1283	-0.1668
3	0.4272	0.2079	0.2622
4	0.2074	-0.0818	-0.1159
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 13 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0495	-0.1645	-0.3487
* 2	0.8142	0.3328	0.4836
3	0.1053	-0.2475	-0.4169
4	0.0310	-0.1506	-0.3734
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 14 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1269	-0.2825	-0.4520
2	0.2012	-0.1776	-0.2534
* 3	0.6625	0.2661	0.3443
4	0.0093	-0.0638	-0.2448
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 15 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0743	-0.2716	-0.5068
* 2	0.7152	0.4644	0.6176
3	0.1517	-0.2416	-0.3689
4	0.0588	-0.2575	-0.5167
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 16 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1146	-0.2750	-0.4525
* 2	0.7957	0.3286	0.4672
3	0.0155	-0.0944	-0.2996
4	0.0743	-0.1971	-0.3679
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 17 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
* 1	0.7152	0.2972	0.3953
2	0.1269	-0.1622	-0.2596
3	0.1115	-0.1476	-0.2447
4	0.0464	-0.1602	-0.3467
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 18 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
* 1	0.3684	-0.0221	-0.0282
2	0.2508	0.0207	0.0282
3	0.1362	-0.1885	-0.2960
4	0.2446	0.1543	0.2112
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 19 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
* 1	0.7988	0.2201	0.3141
2	0.0341	-0.0631	-0.1515
3	0.0464	-0.1368	-0.2961
4	0.1207	-0.1474	-0.2390
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 20 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0681	-0.1464	-0.2806
2	0.0805	-0.3230	-0.5884
3	0.2477	-0.3087	-0.4214
* 4	0.6037	0.4396	0.5579
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 21 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0960	-0.1541	-0.2665
2	0.1455	-0.1718	-0.2652
3	0.2012	-0.2676	-0.3818
* 4	0.5573	0.3301	0.4153
OTHER	.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 22 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0310	-0.2074	-0.5144
2	0.0186	-0.1607	-0.4771
3	0.0526	-0.1528	-0.3177
* 4	0.8978	0.2382	0.4047
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 23 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
* 1	0.6749	0.1955	0.2544
2	0.1950	-0.0946	-0.1360
3	0.0898	-0.1094	-0.1929
4	0.0402	-0.1159	-0.2631
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 24 *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.1517	-0.1069	-0.1632
2	0.3282	-0.1983	-0.2577
3	0.1610	-0.0858	-0.1290
* 4	0.3591	0.2377	0.3051
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 * ITEM 25. *

OPTION	PROPORTION SELECTED	POINT BISERIAL	BISERIAL
1	0.0991	-0.1964	-0.3366
2	0.1207	-0.1637	-0.2655
* 3	0.6687	0.2291	0.2974
4	0.1115	-0.0858	-0.1422
OTHER	0.0000		

VALID CASES: 323
 CASES WITH MISSING VALUES: 0

 *
 * TEST STATISTICS *
 *

TEST MEAN	14.7802
TEST ST. DEV	4.3782
KR-20	0.7470
SEM	2.2020
MISSING FOR ALL ITEMS	0

Note: Item 3 was used for the 0.40 item difficulty conditions while item 17 was used for the 0.70 item difficulty conditions.

Appendix B

C THIS PROGRAM GENERATES THE GROUP ITEM RESPONSES, FOR
 C ITEMS HAVING NALT/4. THE PRODUCT OF THIS PROGRAM IS
 C A DATA FILE HAVING NALT/4 ITEMS CODED 1,2,3, OR 4.
 C THE TOTAL TEST SCORE IS RECORDED IN THIS FILE.

PROGRAM GNPRB

INTEGER NR,NALT,Y,KEY(50),ONE,OPTION(200),
 & ANSWER(1500,50),COUNT,I,TOTAL(1500),
 & SORTA(1500),JJ,NUM,CNTUP,CNTDN,CUT(4),SPACE,
 & CUTOFF(3),GRPAB(1500)

REAL
 THETAR(1500),A(200),B(200),PWRONG(50),R,Z1,Z2,Z3,
 & Z4,XA,XB,XC,XD,SUM,PTOTAL(1500)

DOUBLE PRECISION DSEEDR,DSEED

COUNT=0
 ONE=1
 Y=0

C NALT TOTAL NUMBER OF OPTIONS
 C NR THE TOTAL NUMBER OF SUBJECTS

923 READ(23,923) NALT,NR
 923 FORMAT (I3,1X,I5)
 927 READ(27,927) DSEEDR
 927 FORMAT (F8.1)

DSEED=DSEEDR+100

C GENERATE THE THETA NORMAL(0,1)
 CALL GGNML (DSEEDR,NR,THETAR)

C READ IN THE A AND B VALUES WHICH WERE COMPUTED
 C FROM THE REAL TEST

DO 20 I=1,NALT
 921 READ (21,921) OPTION(I),A(I),B(I)
 921 FORMAT (1X,I2,2X,F7.4,2X,F7.4)
 20 CONTINUE

C IDENTIFY THE CORRECT OPTIONS AND STORE THEM
 C INTO KEY. THIS IS USED LATER TO COMPUTE THE TOTAL
 C SCORE AND TO SELECT THE SUBJECTS WHICH HAVE
 C INCORRECTLY ANSWERED ITEM 1.

```
DO 30 J=1, (NALT/4)
  DO 40 I=1, 4
    Y=Y+1
    IF ((OPTION(Y)).EQ.11) THEN
      KEY(J)=1
    ELSEIF ((OPTION(Y)).EQ.12) THEN
      KEY(J)=2
    ELSEIF ((OPTION(Y)).EQ.13) THEN
      KEY(J)=3
    ELSEIF ((OPTION(Y)).EQ.14) THEN
      KEY(J)=4
    ELSE
      ENDIF
```

```
40 CONTINUE
30 CONTINUE
```

C COMPUTE THE PROBABILITIES OF GETTING EACH OF THE
 C OPTIONS DEPENDING ON THE SUBJECT'S ABILITY.

```
DO 50 H=1, NR
  Y=0
  DO 60 J=1, (NALT/4)
    DO 70 I=1, 4
      Y=Y+1
      IF (I.EQ.1) THEN
        Z1=(1.7*A(Y)*THETAR(H))+B(Y)
      ELSEIF (I.EQ.2) THEN
        Z2=(1.7*A(Y)*THETAR(H))+B(Y)
      ELSEIF (I.EQ.3) THEN
        Z3=(1.7*A(Y)*THETAR(H))+B(Y)
      ELSE
        Z4=(1.7*A(Y)*THETAR(H))+B(Y)
      ENDIF
70 CONTINUE
    SUM=0
    SUM=EXP(Z1)+EXP(Z2)+EXP(Z3)+EXP(Z4)
    XA=EXP(Z1)/SUM
    XB=EXP(Z2)/SUM
    XC=EXP(Z3)/SUM
    XD=EXP(Z4)/SUM
```

C GENERATE A RANDOM NUMBER BETWEEN 0 AND 1, ACTUALLY A
 C REAL NUMBER. THE FOLLOWING ARE USED TO FIND OUT IF
 C THE SUBJECTS WILL GET A 1,2,3, OR 4 ON THE ITEM.

```

      DSEED=DSEED+1
      CALL GGUBS (DSEED, ONE, R)
      XB=XB+XA
      XC=XC+XB
      XD=XD+XC
      IF (R.LE.XA) THEN
        ANSWER(H,J)=1
      ELSEIF (R.GT.XA.AND.R.LE.XB) THEN
        ANSWER(H,J)=2
      ELSEIF (R.GT.XB.AND.R.LE.XC) THEN
        ANSWER(H,J)=3
      ELSE
        ANSWER(H,J)=4
      ENDIF

```

```

60       CONTINUE
50     CONTINUE

```

```

C     INITIALIZE TOTAL
      DO 79 I=1,NR
        TOTAL(I)=0
79     CONTINUE

```

C WRITE OUT ONLY THE SUBJECT'S WHO ANSWERED ITEM 1
 C INCORRECTLY AND COMPUTE THE TOTAL SCORE (IN
 C PERCENTAGE) .

```

      DO 80 I=1,NR
        DO 81 F=1, (NALT/4)
          IF (ANSWER(I,F).EQ.KEY(F)) THEN
            TOTAL(I)=TOTAL(I)+1
          ELSE
            ENDIF
81       CONTINUE
        COUNT=COUNT+1
80     CONTINUE

```

C COUNT INDICATE THE NUMBER OF SUBJECTS WHICH HAVE
 C ANSWERED INCORRECTLY ITEM 1. THIS IS LATER USED
 C IN ANOTHER PROGRAM.

```

      DO 78 I=1,NR
        SORTA(I)=TOTAL(I)
78     CONTINUE

```

```

CALL SORT(NR, SORTA)
SPACE=0
CUT(1)=0

DO 100 I=1,3
  JJ=(NR-CUT(I))/(4-I+1)
  JJ=JJ+CUT(I)
  CNTUP=0
  CNTDN=0
  DO 110 J=1,50
    IF(SORTA(JJ).EQ.SORTA(JJ+J)) THEN
      CNTUP=CNTUP+1
    ELSEIF(SORTA(JJ).EQ.SORTA(JJ-1)) THEN
      CNTDN=CNTDN+1
    ELSE
      CONTINUE
    ENDIF
  110 CONTINUE
  IF(CNTUP.EQ.CNTDN) THEN
    CUTOFF(I)=SORTA(JJ)
    CUT(I+1)=JJ
  ELSEIF(CNTUP.LT.CNTDN) THEN
    CUTOFF(I)=SORTA(JJ)
    CUT(I+1)=JJ+CNTUP
  ELSE
    CUTOFF(I)=SORTA(JJ)-1
    CUT(I+1)=JJ+CNTDN
  ENDIF
100 CONTINUE

CUTOFF(1)=9
CUTOFF(2)=14
CUTOFF(3)=18

DO 120 I=1,NR
  IF(TOTAL(I).LE.CUTOFF(1)) THEN
    GRPAB(I)=1
  ELSEIF((TOTAL(I).GT.CUTOFF(1)).AND.(TOTAL(I)
& .LE.CUTOFF(2))) THEN
    GRPAB(I)=2
  ELSEIF((TOTAL(I).GT.CUTOFF(2)).AND.(TOTAL(I)
& .LE.CUTOFF(3))) THEN
    GRPAB(I)=3
  ELSE
    GRPAB(I)=4
  ENDIF

```

```

        PTOTAL(I)=TOTAL(I)
        PTOTAL(I)=PTOTAL(I)/(NALT/4)*100
        WRITE(25,925)
          (ANSWER(I,J),J=1,(NALT/4)),PTOTAL(I),GRPAB(I)
925      FORMAT(/,5I1,1X,F6.2,1X,I1)
120    CONTINUE

        WRITE(28,928) COUNT
928    FORMAT (I5)

        STOP
        END

```

```

SUBROUTINE SORT(R,ARRAY)
INTEGER R,SPACE,J,ARRAY(R),TEMP

SPACE=R/2.
10    IF(SPACE.LE.0)GO TO 50
        DO 40 I=SPACE+1,R
            J=I-SPACE
20        IF(J.LE.0)GO TO 30
            IF (ARRAY(J).GT.ARRAY(J+SPACE)) THEN
                TEMP=ARRAY(J)
                ARRAY(J)=ARRAY(J+SPACE)
                ARRAY(J+SPACE)=TEMP
                J=J-SPACE
            ELSE
                J=0
            ENDIF
            GO TO 20
30        CONTINUE
40        CONTINUE
        SPACE=SPACE/2.
        GO TO 10
10    CONTINUE
        RETURN
        END

```

Appendix C

Table C: Item difficulty values for the basis, reference, and focal groups

Item		Basis		Reference		Focal	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
1	1	0.1269	0.1366	0.0166	0.1407	0.0239	
	2	0.1115	0.1205	0.0149	0.1215	0.0212	
	3	0.0464	0.0793	0.0118	0.0778	0.0187	
	4*	0.7152	0.6636	0.0212	0.6600	0.0299	
2	1	0.2786	0.2448	0.0202	0.2456	0.0254	
	2	0.1207	0.1617	0.0173	0.1621	0.0238	
	3	0.1889	0.1694	0.0168	0.1725	0.0241	
	4*	0.4118	0.4241	0.0209	0.4198	0.0306	
3	1	0.1207	0.1244	0.0156	0.1244	0.0237	
	2	0.1362	0.1560	0.0171	0.1570	0.0233	
	3	0.1238	0.1365	0.0151	0.1373	0.0221	
	4*	0.6192	0.5831	0.0189	0.5812	0.0266	
4	1	0.0557	0.0995	0.0138	0.1000	0.0185	
	2*	0.5882	0.5567	0.0215	0.5537	0.0275	
	3	0.1207	0.1433	0.0158	0.1450	0.0236	
	4	0.2353	0.2005	0.0181	0.2013	0.0235	
5	1	0.3622	0.3329	0.0236	0.3334	0.0308	
	2	0.2167	0.2106	0.0166	0.2087	0.0272	
	3*	0.3189	0.3227	0.0217	0.3246	0.0279	
	4	0.1022	0.1338	0.0171	0.1332	0.0233	
6	1*	0.5635	0.5399	0.0220	0.5375	0.0332	
	2	0.0433	0.0776	0.0119	0.0784	0.0161	
	3	0.1765	0.1763	0.0161	0.1758	0.0229	
	4	0.2167	0.2062	0.0193	0.2083	0.0277	
7	1	0.2910	0.2580	0.0221	0.2567	0.0298	
	2	0.1517	0.1735	0.0174	0.1763	0.0250	
	3*	0.4830	0.4663	0.0221	0.4638	0.0338	
	4	0.0743	0.1022	0.0141	0.1032	0.0212	
8	1	0.1424	0.1615	0.0160	0.1656	0.0225	
	2*	0.6904	0.6377	0.0214	0.6319	0.0302	
	3	0.1084	0.1181	0.0129	0.1194	0.0205	
	4	0.0588	0.0827	0.0108	0.0831	0.0164	

Table C: Item difficulty values for the basis, reference, and focal groups (continued).

Item		Basis		Reference		Focal	
			Mean	S.D.		Mean	S.D.
9	1	0.1703	0.1862	0.0179	0.1875	0.0251	
	2*	0.4396	0.4177	0.0243	0.4170	0.0336	
	3	0.2446	0.2430	0.0216	0.2406	0.0276	
	4	0.1455	0.1531	0.0177	0.1549	0.0257	
10	1	0.0805	0.1213	0.0148	0.1230	0.0195	
	2	0.0960	0.1103	0.0147	0.1101	0.0209	
	3	0.1672	0.1584	0.0169	0.1597	0.0214	
	4*	0.6563	0.6101	0.0214	0.6072	0.0295	
11	1	0.1796	0.1694	0.0180	0.1689	0.0240	
	2	0.2724	0.2684	0.0221	0.2663	0.0290	
	3*	0.4830	0.4642	0.0238	0.4658	0.0350	
	4	0.0650	0.0981	0.0138	0.0990	0.0181	
12	1	0.0650	0.1097	0.0147	0.1102	0.0189	
	2	0.0991	0.1267	0.0154	0.1284	0.0201	
	3*	0.5666	0.5378	0.0217	0.5375	0.0334	
	4	0.2693	0.2258	0.0192	0.2238	0.0273	
13	1	0.0372	0.0776	0.0113	0.0782	0.0173	
	2*	0.3282	0.3047	0.0223	0.3070	0.0325	
	3	0.4272	0.4097	0.0235	0.4064	0.0350	
	4	0.2074	0.2079	0.0172	0.2084	0.0267	
14	1	0.0495	0.0711	0.0118	0.0706	0.0153	
	2*	0.8142	0.7468	0.0195	0.7435	0.0363	
	3	0.1053	0.1230	0.0136	0.1240	0.0212	
	4	0.0310	0.0592	0.0108	0.0619	0.0147	
15	1	0.1269	0.1564	0.0147	0.1547	0.0234	
	2	0.2012	0.1819	0.0181	0.1822	0.0235	
	3*	0.6625	0.6245	0.0219	0.6247	0.0325	
	4	0.0093	0.0372	0.0088	0.0384	0.0118	
16	1	0.0743	0.1080	0.0124	0.1074	0.0197	
	2*	0.7152	0.6546	0.0202	0.6534	0.0308	
	3	0.1517	0.1436	0.0147	0.1457	0.0219	
	4	0.0588	0.0939	0.0129	0.0935	0.0191	
17	1	0.1146	0.1310	0.0145	0.1327	0.0217	
	2*	0.7957	0.7339	0.0181	0.7320	0.0250	
	3	0.0155	0.0434	0.0083	0.0444	0.0116	
	4	0.0743	0.0918	0.0119	0.0909	0.0163	

Table C: Item difficulty values for the basis, reference, and focal groups (continued).

Item	Basis	Reference		Focal		
		Mean	S.D.	Mean	S.D.	
18	1*	0.3684	0.3400	0.0198	0.3401	0.0286
	2	0.2508	0.2412	0.0183	0.2388	0.0292
	3	0.1362	0.1708	0.0173	0.1720	0.0227
	4	0.2446	0.2480	0.0186	0.2492	0.0293
19	1*	0.7988	0.7553	0.0193	0.7563	0.0283
	2	0.0341	0.0542	0.0101	0.0538	0.0147
	3	0.0464	0.0731	0.0118	0.0712	0.0160
	4	0.1207	0.1173	0.0155	0.1186	0.0236
20	1	0.0681	0.0874	0.0124	0.0888	0.0181
	2	0.0805	0.1232	0.0151	0.1232	0.0233
	3	0.2477	0.2227	0.0164	0.2256	0.0268
	4*	0.6037	0.5666	0.0207	0.5624	0.0293
21	1	0.0960	0.1144	0.0143	0.1151	0.0193
	2	0.1455	0.1522	0.0155	0.1518	0.0210
	3	0.2012	0.2052	0.0189	0.2070	0.0254
	4*	0.5573	0.5282	0.0233	0.5260	0.0331
22	1	0.0310	0.0593	0.0094	0.0590	0.0144
	2	0.0186	0.0419	0.0079	0.0436	0.0126
	3	0.0526	0.0596	0.0101	0.0599	0.0154
	4*	0.8978	0.8392	0.0147	0.8375	0.0237
23	1*	0.6749	0.6386	0.0216	0.6377	0.0271
	2	0.1950	0.1755	0.0171	0.1764	0.0219
	3	0.0898	0.1103	0.0159	0.1090	0.0206
	4	0.0402	0.0755	0.0111	0.0769	0.0176
24	1	0.1517	0.1648	0.0164	0.1636	0.0224
	2	0.3282	0.3077	0.0189	0.3077	0.0264
	3	0.1610	0.1706	0.0154	0.1715	0.0229
	4*	0.3591	0.3569	0.0208	0.3572	0.0286
25	1	0.0991	0.1217	0.0138	0.1226	0.0230
	2	0.1207	0.1328	0.0162	0.1340	0.0205
	3*	0.6687	0.6274	0.0227	0.6243	0.0325
	4	0.1115	0.1182	0.0135	0.1191	0.0217

Appendix D

Table D: Item biserial correlations for the basis, reference, and focal groups.

Item		Basis	Reference		Focal	
			Mean	S.D.	Mean	S.D.
1	1	-0.2596	-0.4354	0.0596	-0.4359	0.0948
	2	-0.2447	-0.3935	0.0658	-0.3862	0.0891
	3	-0.3467	-0.4214	0.0757	-0.4116	0.1085
	4*	0.3953	0.5603	0.0506	0.5532	0.0667
2	1	-0.2711	-0.3457	0.0467	-0.3360	0.0725
	2	-0.5250	-0.5410	0.0594	-0.5425	0.0790
	3	-0.1486	-0.2073	0.0676	-0.2060	0.0865
	4*	0.6082	0.6710	0.0373	0.6683	0.0533
3	1	-0.1458	-0.2195	0.0626	-0.2146	0.0898
	2	-0.3765	-0.4466	0.0647	-0.4508	0.0962
	3	-0.2670	-0.3450	0.0668	-0.3451	0.0903
	4*	0.3046	0.4886	0.0543	0.4907	0.0709
4	1	-0.4595	-0.4650	0.0600	-0.4724	0.0905
	2*	0.4325	0.5844	0.0444	0.5881	0.0638
	3	-0.3516	-0.3703	0.0575	-0.3728	0.0770
	4	-0.1674	-0.2577	0.0629	-0.2551	0.0905
5	1	-0.1986	-0.2579	0.0573	-0.2620	0.0803
	2	-0.0665	-0.0792	0.0591	-0.0870	0.0940
	3*	0.2823	0.3792	0.0493	0.3855	0.0672
	4	-0.2663	-0.2244	0.0755	-0.2205	0.1015
6	1*	0.5399	0.6602	0.0405	0.6668	0.0634
	2	-0.3497	-0.3314	0.0849	-0.3498	0.1138
	3	-0.3240	-0.3730	0.0585	-0.3677	0.0860
	4	-0.3289	-0.4104	0.0591	-0.4120	0.0827
7	1	-0.2214	-0.3061	0.0569	-0.2998	0.0855
	2	-0.4010	-0.4408	0.0589	-0.4454	0.0871
	3*	0.3975	0.5502	0.0449	0.5530	0.0552
	4	-0.2265	-0.1973	0.0741	-0.2012	0.1061
8	1	-0.4142	-0.5490	0.0570	-0.5518	0.0867
	2*	0.4496	0.6335	0.0441	0.6415	0.0684
	3	-0.3019	-0.3607	0.0695	-0.3665	0.0958
	4	-0.2434	-0.2927	0.0813	-0.2897	0.1100

Table D: Item biserial correlations for the basis, reference, and focal groups (continued).

Item		Basis		Reference		Focal	
			Mean	S.D.		Mean	S.D.
9	1	-0.2929	-0.3372	0.0591	-0.3394	0.0873	
	2*	0.3169	0.4280	0.0521	0.4398	0.0653	
	3	-0.2750	-0.3285	0.0584	-0.3395	0.0762	
	4	0.0898	0.0629	0.0612	0.0594	0.0758	
10	1	-0.5553	-0.5659	0.0529	-0.5642	0.0744	
	2	-0.3869	-0.4191	0.0662	-0.4186	0.0948	
	3	-0.3647	-0.4421	0.0604	-0.4331	0.0886	
	4*	0.5398	0.7059	0.0397	0.7024	0.0544	
11	1	-0.0897	-0.1348	0.0693	-0.1328	0.0968	
	2	-0.3804	-0.4810	0.0550	-0.4843	0.0787	
	3*	0.3649	0.5262	0.0459	0.5294	0.0623	
	4	-0.3264	-0.2682	0.0717	-0.2705	0.1034	
12	1	-0.4482	-0.4713	0.0645	-0.4761	0.0943	
	2	-0.3285	-0.3630	0.0602	-0.3632	0.0860	
	3*	0.3683	0.5256	0.0433	0.5336	0.0652	
	4	-0.1461	-0.2056	0.0565	-0.2104	0.0843	
13	1	-0.1513	-0.1600	0.0750	-0.1526	0.1143	
	2*	-0.1668	-0.2228	0.0523	-0.2283	0.0765	
	3	0.2622	0.3611	0.0533	0.3600	0.0766	
	4	-0.1159	-0.1527	0.0609	-0.1465	0.0999	
14	1	-0.3487	-0.4678	0.0769	-0.4651	0.1102	
	2*	0.4836	0.6548	0.0459	0.6548	0.0637	
	3	-0.4169	-0.5275	0.0563	-0.5291	0.0809	
	4	-0.3734	-0.3876	0.0750	-0.3755	0.1142	
15	1	-0.4520	-0.5327	0.0618	-0.5302	0.0835	
	2	-0.2534	-0.3745	0.0569	-0.3737	0.0685	
	3*	0.3443	0.5566	0.0556	0.5557	0.0716	
	4	-0.2448	-0.1821	0.1123	-0.1900	0.1407	
16	1	-0.5068	-0.5614	0.0583	-0.5609	0.0827	
	2*	0.6176	0.7647	0.0387	0.7651	0.0544	
	3	-0.3689	-0.4431	0.0572	-0.4490	0.0854	
	4	-0.5167	-0.5072	0.0667	-0.4996	0.0921	
17	1	-0.4525	-0.6121	0.0595	-0.6232	0.0858	
	2*	0.4672	0.6625	0.0459	0.6700	0.0613	
	3	-0.2996	-0.3116	0.0952	-0.3078	0.1326	
	4	-0.3679	-0.4298	0.0805	-0.4296	0.1118	

Table D: Item biserial correlations for the basis, reference, and focal groups (continued).

Item		Basis		Reference		Focal	
			Mean	S.D.		Mean	S.D.
18	1*	-0.0282	-0.0306	0.0536	-0.0352	0.0741	
	2	0.0282	0.0491	0.0593	0.0526	0.0857	
	3	-0.2960	-0.3098	0.0694	-0.3079	0.1075	
	4	0.2112	0.2370	0.0582	0.2381	0.0757	
19	1*	0.3141	0.4754	0.0544	0.4837	0.0737	
	2	-0.1515	-0.2350	0.0973	-0.2575	0.1265	
	3	-0.2961	-0.3739	0.0709	-0.3659	0.0998	
	4	-0.2390	-0.3633	0.0723	-0.3722	0.1054	
20	1	-0.2806	-0.2861	0.0616	-0.2824	0.0963	
	2	-0.5884	-0.5730	0.0660	-0.5676	0.0871	
	3	-0.4214	-0.5052	0.0546	-0.5056	0.0765	
	4*	0.5579	0.7224	0.0397	0.7211	0.0532	
21	1	-0.2665	-0.3012	0.0700	-0.3037	0.0963	
	2	-0.2652	-0.3132	0.0634	-0.3144	0.0826	
	3	-0.3818	-0.4797	0.0506	-0.4779	0.0742	
	4*	0.4153	0.5869	0.0442	0.5903	0.0618	
22	1	-0.5144	-0.5964	0.0666	-0.6003	0.1004	
	2	-0.4771	-0.5348	0.0856	-0.5372	0.1281	
	3	-0.3177	-0.4610	0.0624	-0.4772	0.0911	
	4*	0.4047	0.6274	0.0538	0.6376	0.0733	
23	1*	0.2544	0.3917	0.0510	0.3844	0.0702	
	2	-0.1360	-0.2253	0.0586	-0.2143	0.0857	
	3	-0.1929	-0.2524	0.0665	-0.2476	0.1071	
	4	-0.2631	-0.2845	0.0849	-0.2867	0.1186	
24	1	-0.1632	-0.1671	0.0662	-0.1819	0.0986	
	2	-0.2577	-0.3415	0.0584	-0.3387	0.0747	
	3	-0.1290	-0.1340	0.0651	-0.1312	0.0825	
	4*	0.3051	0.4267	0.0436	0.4343	0.0668	
25	1	-0.3366	-0.4156	0.0728	-0.4213	0.0975	
	2	-0.2655	-0.3727	0.0684	-0.3680	0.0901	
	3*	0.2974	0.4617	0.0533	0.4629	0.0695	
	4	-0.1422	-0.1744	0.0725	-0.1692	0.1044	