

1-1-1995

Detection of differential item functioning in multiple language groups using item response theory and logistic regression procedures.

Anil Kanjee
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Kanjee, Anil, "Detection of differential item functioning in multiple language groups using item response theory and logistic regression procedures." (1995). *Doctoral Dissertations 1896 - February 2014*. 5194. https://scholarworks.umass.edu/dissertations_1/5194

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST



312066011267577

DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN MULTIPLE
LANGUAGE GROUPS USING ITEM RESPONSE THEORY AND
LOGISTIC REGRESSION PROCEDURES

A Dissertation

by

ANIL KANJEE

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

February 1995

School of Education

© Copyright by Anil Kanjee, 1995
All Rights Reserved

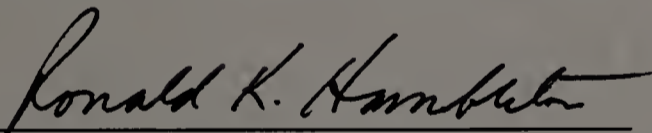
DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN MULTIPLE
LANGUAGE GROUPS USING ITEM RESPONSE THEORY AND
LOGISTIC REGRESSION PROCEDURES

A Dissertation

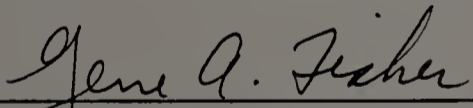
by

ANIL KANJEE

Approved as to style and content by:



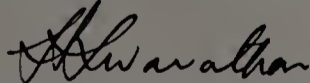
Ronald K. Hambleton, Chair



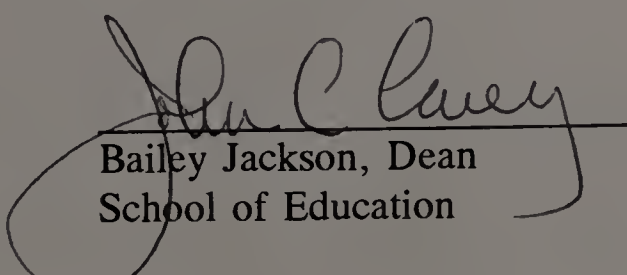
Gene A. Fisher, Member



Jane Rogers, Member



Hariharan Swaminathan, Member



Bailey Jackson, Dean
School of Education

This study is dedicated to my parents
Soma and Indira Kanjee
and all other
South African parents
who
like them
have committed themselves
to the education
of our
youth
for advancing
the struggle
of
our people

ACKNOWLEDGMENTS

There have been people, both in the U.S. and in South Africa, who have contributed in various different ways, politically, economically, socially, and psychologically, to my successfully completing my educational mission in the U.S., including this study. For this, I am eternally gratefully and hope to continue to live my life in this spirit of friendship. It has been an extremely long and often difficult five and half years away from home, more so because of all the historical changes that have affected our people in this period.

While it is not possible to acknowledge everybody, and all that has been done for me, a few people need to be mentioned. My entire community has always provided me with unwavering support and encouragement, especially my Mom and Dad, my brothers, Vijesh and Nayen, and my friend, Gordon. It is my hope that on my return I will be able to pay back the favor.

Both students and faculty in my program have created an academic environment that for me, made it easier and more enjoyable to acquire the relevant knowledge and skill I required. Specifically, my advisor, Ron Hambleton, who not only provided valuable advise, training and academic assistance, but also opened up his home and family to me, and I always knew that I could rely on Ron if ever there was a need. Ron has always been confident in my ability and work, and assisted whenever possible. Swami, whom I consider as my second advisor, has always supported me in my work and has been a source of inspiration in how to teach the 'dreaded' statistics course. Both my advisors certainly made a significant difference (at the .001 level at least) during my stay. I would like to

thank Jane, who always found time for me whenever it was needed, for being on my committee, for discussing and clarifying various ideas related to this study, and especially for providing the statistical and programming expertise that made this study a success. I also would like to thank Dr. Gene Fisher for taking the time to serve on my committee, and for the assistance and guidance provided.

From my first day in the program, Peg has been providing assistance and information, from making accommodation arrangements to typing guidelines. Many thanks to my all my classmates especially Pankaja, Mohamed and Kathy. Also, the advice, companionship and assistance from my fellow South African comrade and friend, Mohapi, has been always available, and I hope that this spirit prevails for the future to the benefit of our people. Special thanks goes to Else Hambleton, for opening up her home to me as well for the political conversations we've had about South Africa.

I would also like to thank the Desirée, Craig, Keegan and Kyle, who were both my friends and family in Amherst. I am especially indebted to Linda Schade, who not only provided valuable editing assistance for this study, but also provided the support and encouragement that could only come from a loving and caring partner. Finally there are many friends and comrades, without whom my stay here would be meaningless, who should be noted: Razack Karriem, Rashid Ahmed, Shenid Bhayroo, Jennifer Cannon, Anilla Cherian, Hamid Elhasnaouie, Irieza Fortune, Rahel Gottlieb, Emily Katz, Eyad Kishawi, Verna Lalbeharie, Lusani Madziuhandila, Mostafa Mouhieddien, Moketsi Mosola, Morongoe Ntlodibe, Bava Pillay, Sivan Pillay, Shajila Singh, Silaine Souza, Jim Statman, and Roberta Uno.

ABSTRACT

DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN MULTIPLE LANGUAGE GROUPS USING ITEM RESPONSE THEORY AND LOGISTIC REGRESSION PROCEDURES

FEBRUARY 1995

ANIL KANJEE, B.Sc., UNIVERSITY OF DURBAN-WESTVILLE

H.D.E. (PG), UNIVERSITY OF CAPE TOWN

B.A. (HONORS), UNIVERSITY OF THE WESTERN CAPE

M. Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

Small sample sizes are a problem facing many assessment practitioners when conducting differential item functioning (DIF) studies. Although one of the main reasons for conducting DIF studies is to ensure that assessment instruments are not biased against minority groups, when sample sizes are small, it is common practice to by-pass DIF studies altogether. This is because popular DIF detection procedures with small samples often leads to unreliable and invalid results.

A second problem is that most DIF procedures are only applicable to two-group comparisons. When three or more groups are compared, the use of two-group comparisons are problematic since many comparisons may be required, and it is not clear whether items identified as DIF between two groups also function differentially for other groups. Also, type I error rates are inflated.

The purpose of this study was to extend two promising DIF detection

procedures, the pseudo-IRT and the logistic regression (LR) procedures, in order to address the problem of small sample size by simultaneously detecting DIF in multiple groups.

Item response data were simulated for three groups with 10% of the 60 items simulated as DIF in groups 2 and 3. Three estimation procedures for the pseudo-IRT procedure, and two for the LR procedure, were used. Sample sizes as well as the mean ability distributions were also varied.

Results indicated that both the pseudo-IRT and LR procedures could successfully be extended to simultaneously detect DIF in multiple groups when sample sizes were large. When sample sizes decreased, the number of DIF items detected also decreased. Varying the mean ability distribution of one of the groups also significantly affected the results. The LR procedure detected slightly more items than the pseudo-IRT procedures. With sample sizes of 100, the detection rate was low for both procedures. The use of simultaneous procedures did enable some analysis of groups with small samples, though power to detect DIF was low.

The significance of this study is that comparisons with samples as low as 100 can provide useful information. However, further research is required to improve the methodology for DIF studies with small samples.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGMENTS	v
ABSTRACT	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiii
CHAPTER	
1. INTRODUCTION	1
1.1 Background	1
1.2 History of Assessment Practices	6
1.3 Cross-Cultural/National Studies	8
1.4 Assessment Practices in South Africa	10
1.4.1 Assessment and Education in South Africa	11
1.4.2 Educational Assessment and Translation	16
1.4.3 Possible Problems and Obstacles	17
1.5 Purposes of the Study	20
1.6 Organization of the Thesis	21
2. REVIEW OF THE LITERATURE	24
2.1 Introduction	24
2.2 Reasons for Translating Instruments	25
2.3 Equivalence and DIF in Cross- Cultural/Language Comparisons	27
2.4 Judgmental Designs for Assessing Equivalence	28
2.4.1 Forward-Translation Designs	29
2.4.2 Back-Translation Designs	31
2.5 Statistical Methods of Assessing Equivalence	33
2.5.1 Statistical Designs	33
2.5.2 Statistical Techniques to Detect DIF	36

2.6	Questions That This Study Addresses	52
3.	DETECTION OF DIF IN MULTIPLE GROUPS	56
3.1	Introduction	56
3.2	Pseudo-IRT Estimation Procedures	58
	3.2.1 Pseudo-IRT Procedure 1 (PIRT 1)	58
	3.2.2 Pseudo-IRT Procedure 2 (PIRT 2)	59
	3.2.3 Pseudo-IRT Procedure 3 (PIRT 3)	60
3.3	Logistic Regression Estimation Procedures	61
3.4	Purpose and Research Questions	63
3.5	Method	64
	3.5.1 Description of Data	64
	3.5.2 Procedure	66
3.6	Results	68
3.7	Discussion	76
4.	THE EFFECT OF SAMPLE SIZE AND ABILITY DISTRIBUTION ON THE DETECTION OF DIF IN MULTIPLE GROUPS	89
4.1	Purpose of the Investigation	89
4.2	Research Design	93
	4.2.1 Sample Size	94
	4.2.2 Ability Distributions	95
	4.2.3 Description of DIF Procedures	95
	4.2.4 Evaluation Criteria	95
4.3	Results and Discussion	95
	4.3.1 Effect of Sample Size	96
	4.3.2 Effect of Different Mean Ability Distribution	102
	4.3.3 Summary	107
5.	SUMMARY AND CONCLUSIONS	120
5.1	Summary	120
5.2	Significance of the Findings	125
5.3	Implications and Recommendations	127
5.4	Shortcomings	130
5.5	Directions for Further Research	131

APPENDICES

A. PERCENTAGE OF PEOPLE AND LANGUAGES SPOKEN IN SOUTH AFRICA 133

B. STUDENT-TEACHER AND -CLASSROOM RATIOS BY "RACE" GROUP 135

REFERENCES 137

LIST OF TABLES

Table		
3.1	Item Numbers and a- and b-values for Uniform and Non-Uniform DIF Items	81
3.2	Percentage of DIF Items Detected (over 20 replications) for Pseudo-IRT Procedures 1, 2 and 3	82
3.3	Percentage of DIF Items Detected (over 20 replications) for LR procedures 1 and 2	83
3.4	Percentage of Item Types Detected Using Pseudo-IRT Procedures	84
3.5	Percentage of Item Types Detected Using Logistic Regression Procedures	85
4.1	Group Sample Sizes, Mean Abilities and Standard Deviations	110
4.2	Effect of Sample Size on the Percentage of DIF Items (over 20 replications) Detected for Using the Pseudo-IRT Procedures 1, 2 and 3	111
4.3	Effect of Sample Size on the Percentage of DIF Items (over 20 years) Detected Using the LR Procedure 1 and 2	112
4.4	Effect of Sample Size on the Percentage of DIF Items Detected Using Pseudo-IRT Procedures	113
4.5	Effect of Sample Size on the Percentage of DIF Items Detected Using Logistic Regression Procedures	114
4.6	Effect of Sample Size and Ability Distribution on the Percentage of DIF Items Detected Using Pseudo-IRT Procedures	115
4.7	Effect of Sample Size and Ability Distribution on the Percentage of DIF Items Detected Using Logistic Regression Procedures	116
4.8	Effect of Varying Ability Distributions on the Percentage of Item Types Detected Using Pseudo-IRT Procedures	117
4.9	Effect of Varying Ability Distributions on the Percentage of Item Types Detected Using Logistic Regression Procedures	118

LIST OF FIGURES

Figure

1.1	Education funding in South Africa for 1992-1993	23
1.2	Illiteracy rates in South Africa for 1993	23
2.1	An ICC and ability distributions for two groups of examinees	54
2.2	ICCs showing uniform DIF between target and reference groups	54
2.3	ICCs showing non-uniform DIF between target and reference groups	55
2.4	Observed and expected proportion correct as a function of ability for an item with large negative differences	55
3.1	Uniform DIF item with area = 0.4	86
3.2	Uniform DIF item with area = 0.6	86
3.3	Uniform DIF item with area = 0.8	87
3.4	Low difficulty non-uniform DIF item (area = 0.8)	87
3.5	Moderate difficulty non-uniform DIF item (area = 0.8)	88
3.6	High difficulty non-uniform DIF item (area = 0.8)	88
4.1	An example showing two ICCs that are far apart but when their associated ability distributions are considered, the differences in ICCs affects no one	119

CHAPTER 1

INTRODUCTION

1.1 Background

Recent changes in the political structure in South Africa demand the eradication of the racist apartheid education structure and the creation of a progressive anti-sexist, anti-racist education system. This requires the integration of 14 different education departments and 11 different language and cultural groups into a single, centralized system. Unfortunately, past apartheid practices were inherently biased against the indigenous cultural and language groups (the majority of the population), and biased towards the English and Afrikaans language and cultural groups (refer to Appendices A and B), making the integration extremely difficult.

A central aspect of the new system will be the definition, development and implementation of a democratic and progressive assessment structure free of the biases and prejudices of the old system. This will require the development of appropriate standardized assessment instruments that are both inclusive of and fair to all language groups by assessment practitioners¹. In order to maximize inclusion, examinees must be able to take tests in the language of their choice (first or best language), a process which requires that testing instruments be translated and/or adapted. Once this fundamental barrier is overcome, the next step is to ensure that

¹This is meant to be an all inclusive term to refer to all those involved in the development and use of measuring instruments, including educators, measurement specialists, test developers and administrators.

scores derived from the different versions of the translated/adapted instruments are equivalent.

Assessment devices developed in different languages must be equivalent if direct comparisons of the performance between two (or more) groups are to be made. Drasgow (1984) notes that measurement equivalence exists when the relations between observed test scores and the latent attribute measured by the tests are identical across subpopulations (p. 134). In situations where different nationalities and/or cultural groups are compared, factors like cultural and ethnic differences, in addition to language differences, need to be considered as well. Given this, assessment instruments are said to exhibit measurement equivalence when individuals who are equal on the trait measured by an instrument, but who come from different cultural and linguistic groups, have the same observed scores. Unless the scores from different assessment instruments are equivalent, it is uncertain whether the scores represent valid group differences and similarities in performance or merely measurement artifacts (Ellis, 1989).

Rather than defining measurement equivalence in terms of observed scores, measurement equivalence can be defined in terms of the probability of responding correctly to an item. This notion lends itself to item analysis and item bias. One way of achieving equivalent scores is to develop items that are not biased in favor of or against any one or more groups. In this respect, analyses are conducted to identify any item bias, and either eliminate or revise these items. However, recently, the term 'item bias' has been replaced by the less value laden term differential item functioning (DIF) in acknowledgement of the fact that statistical

methods cannot detect bias as such, only evidence of differential performance (Holland & Thayer, 1988). Thus an item that exhibits DIF may or may not be biased for or against any group. Hambleton, Swaminathan and Rogers (1991) note that the accepted definition of DIF by psychometricians is that an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right (p. 110).

Most studies regarding the detection of DIF have been conducted on two groups only (Hambleton & Rogers, 1989; Holland & Thayer, 1988; Linn & Harnisch, 1981; Swaminathan & Rogers, 1990). Usually the reference group is compared to the target group². Typically (in the U. S), the group with the larger sample size (usually White examinees) is used as the reference group, while the target group(s) selected consists of examinees from ethnic minorities, that is, Asian, Black, Hispanic, and Native American (Zieky, 1993). Items that are flagged as DIF, are either removed from the testing instrument (or from the analysis) or revised and kept in the instrument (Ellis, 1991). When more than two groups are studied, the solution has been to conduct multiple pairwise comparisons which utilize the same two-group comparison techniques (Kim, Cohen, & Park, 1993). That is, a single group (usually the largest group) is utilized as the reference against which the performance of all other groups in the study are compared.

One of the biggest problems with DIF studies is that the sample sizes of target groups found in practice are typically small (Linn & Harnisch, 1981). This is

²Target group refers to the particular group of interest, while the reference group refers to the group against which the standard of comparison for performance is made.

to be expected since the very nature of DIF depends on the analysis of data from minority groups (Zieky, 1993). The consequence of conducting analysis on groups with small sample sizes is that the resulting DIF statistics are not very stable. Linn (1993) noted that at present there is no clear definition of what constitutes an "adequate" sample size for DIF studies. Also, what constitutes an "adequate sample size" depends as much on the specific statistical techniques applied to detect DIF (for example, item response theory based techniques require larger sample sizes than the Mantel-Haenszel techniques) as it does on the purpose of conducting DIF studies (for example, greater rigor is required for the development of a highly selective testing instruments than for the development of an attitude survey for a research study or the development of a classroom achievement test).

In many situations, DIF analysis for groups that consist of small samples is simply not done. The argument is that analysis conducted on small samples results in high sampling errors which increase the number of false positive and false negative errors. While it is true that information derived from studies with small sample sizes are likely to be unreliable, the argument that these studies are of limited value is not acceptable. This practice is especially common when multiple groups are compared, where it is simply reported that analyses of group X or Y could not be conducted due to the lack of an adequate sample size, and the instrument is accepted without further question. For example, many standardized assessment instruments do not report any information on "Native American Indians". Even though "Native American Indians" constitute a small percentage of the U.S.

population, this should not be used as an excuse for excluding them from any analysis simply because a large enough sample size was not collected.

Clearly, an important problem for assessment practitioners is that of sample size. One possible solution for increasing sample sizes is to make use of all the data available. That is, data from different groups being compared could be combined (Linn & Harnisch, 1981). This solution is especially relevant when multiple groups are compared. In this context, the likelihood of creating larger sample sizes is directly proportional to the number of groups in the study. That is, the more groups (which themselves may consist of small samples) that are compared, the greater the probability of creating larger sample sizes. However, data on specific techniques to do this are not readily available.

The aim of this study was to investigate the effect of small sample size on two methods of detecting DIF when three or more groups are compared. Two procedures, one based on item response theory (IRT) and one based on logistic regression (LR) were extended to accommodate multiple group comparisons when sample sizes of one or more of the groups are small. Specifically, this study investigated the detection of DIF when groups from multiple language backgrounds are compared. For example, a requirement in the new education system proposed in South Africa is that students be allowed to choose from any one of the 11 languages (predominantly at the regional level) as their medium of instruction. Thus, any national standardized test needs to be developed in at least 11 languages. In this process, DIF studies are crucial to ensure that all the instruments in the different languages are equivalent.

An additional point needs to be made with regard to the detection of DIF. That is, measurement practitioners and specialists need to acknowledge and recognize that while some of the root causes of DIF are primarily psychometric in nature, others are distinctly historical. This is especially relevant given the use (abuse) of assessment practices in South Africa. The responsible development, application and use of assessment techniques must not be solely based on psychometric and statistical considerations only, but must be informed by the historical context in which assessment takes place. This is especially true with reference to members of minority and/or traditionally oppressed groups. In the next section, a brief history of assessment practices and philosophy is presented. Also, the uses of cross-cultural research is discussed as this is the setting in which the comparison of multiple groups with different languages are most likely to occur.

1.2 History of Assessment Practices

Historically, assessment practices have been primarily used for selection, classification and differentiation between individuals and groups, from the Chinese emperors to Binet, Burt and Jensen to current day specialists (Popham, 1990; Sax, 1989). An unfortunate and racist period in this history has seen the use of assessment to promote the ideology of white superiority and black inferiority (Apple, 1989; Bulhan, 1981; Gould, 1981; Norris, 1990). Gould (1981) notes that primarily, this ideology has had two sources of support: craniometry (measurement of the skull), and the IQ controversy, as manifested by certain styles of testing (Gould, 1981). Norris (1990) articulates this very clearly:

the measurement of individual differences provided a technology of classification for the 'objective' allocation or exclusion of individuals to or from roles, treatments or forms of institutional provision (p. 27).

While such practices have been declared illegal in many countries, there still exists a strong body of opinion and practice which regards assessment as a means of maintaining social control and promoting the status quo, based mainly on race and ethnicity (Bulhan, 1981; Apple, 1989). The education system in South Africa is the quintessential example (Kanjee, 1993a). Popham (1990) portrays this philosophy quite clearly:

the most insidious form of bias in educational testing is ethnic bias, because testing practices that are ethnically biased serve to stifle the attainment of individuals who have often already been served up more than their share of social inequalities. There are so many factors which operate to oppress people from minority groups that it constitutes a major educational tragedy when the progress of minority youngsters is stifled because of bias in testing (p. 178).

The crucial point is that measurement specialists and assessment practitioners need to be not only aware, but sensitive as well, to this history. It needs to be noted that the consequences of this history are still being experienced by many peoples, socially, politically, economically and academically - especially by groups that have been historically discriminated against. These factors must be seriously considered whenever assessment instruments are developed, administered, interpreted and translated. Failure to attend to this concern could result in the construction of biased tests, the misinterpretation of test results and the complete misperception and misunderstanding of different (ethnic and cultural) groups or nations.

1.3 Cross-Cultural/National Studies

Cross-cultural studies emerged primarily from the disciplinary traditions of psychology. As a result, cross-cultural studies have tended to emphasize psychological factors (anxiety, depression) as opposed to educational factors (learning styles, curricula, teaching methodology) when comparing different cultures and/or language groups (Lonner & Berry, 1986; Triandis & Berry, 1980). The consequences of this "psychological bias" is that many of the methods currently used in cross-cultural/national educational assessment were directly appropriated from this field (Hui & Triandis, 1985; Irvine & Carroll, 1980). As a result, methods for cross-national comparisons specific to education (or pedagogy), and the unique problems to this field were not readily developed.

A good example of this is that currently no internationally accepted standards for conducting cross-cultural/language educational assessment research exist (Hambleton, 1993), even though cross-cultural studies have been conducted for over fifty years. Nevertheless, recent theoretical and technological advances in the area of educational assessment have provided a great number of new and innovative techniques that could prove relevant and useful in the context of cross-national educational assessment (Hambleton & Swaminathan, 1985; Holland & Thayer, 1988; Linn, 1989; Nitko, 1989).

Another problem that emerged directly from the psychological tradition is that historically, cross-cultural studies have not been conducted in an atmosphere of reciprocity, where all cultures involved in the study stood to benefit. Rather, information from foreign and [especially] unknown cultures and nations was merely

appropriated (Bulhan, 1981). This is precisely what Jahoda, a noted psychologist, referred to when he noted that:

the invasion by foreign researchers is apt to be viewed as a more or less subtle form of exploitation. In other words, psychology stands accused of gaining advantage from developing countries without providing tangible benefits in return. Those of us who have carried out cross-cultural studies in the past, and are honest with themselves can hardly deny that there is some substance in this charge (cited in Bulhan, 1981, p. 27).

In this respect, fundamental changes in the manner in which cross-national/cultural studies are conducted must occur. That is, the goal of these studies should be for the exchange of information and ideas to the mutual benefit of all respective nationalities and cultures concerned. It is within this spirit that all cross-national studies should be conducted, especially when comparing educational systems and practices, something that is so basic yet so crucial to the development and advancement of all nations and cultures. To this end, the work of the International Association for the Evaluation of Educational Achievement (IEA) over the past 30 years needs to be acknowledged (Wolf, 1992).

An apparent and indisputable example that clearly demonstrates the abuse of cross-cultural studies can be found in the education system in South Africa. Some of the consequences of the racist assessment practices, that were so prevalent and dominant in the early history of assessment, on the peoples of South Africa are well documented (Mathonsi, 1988; Nkomo, 1990; Wolpe, 1992). In the next section, the role of education in South Africa is discussed: (1) to demonstrate a case in point, (2) as this very theme is the current focus of debates about changes in the education system and the nation as a whole, and (3) the methods and designs used in this study

are especially applicable to assessment practices in South Africa with 11 national language and/or cultural groups.

1.4 Assessment Practices in South Africa

The primary focus of this section relates to some of the concerns and issues involved in the current debate for the establishment of a new education system in a non-racial, non-sexist, non-exploitative South African society. The first part discusses the use of the educational system, specifically assessment practices, as means of social control towards the maintenance of the status quo. It is argued that in order to implement a new system of assessment based on the principle of equality for all racial, ethnic and language groups, one has to understand the role of assessment in apartheid education and the consequences of these practices on the lives of both Black and White South Africans. In the second part, the focus is on the role and use of translation practices in educational assessment in South Africa.

The political changes in South Africa, introduced after 1991, have had consequences throughout the entire social and economic structure of South African society. Nowhere have these consequences had a greater impact than in the educational arena. Current debates revolve largely around methods of integrating the different ethnic education departments, centralized versus federal administration systems, teacher training, curriculum content, financial issues, language policy and educational standards. (Nkomo, 1991; Wolpe, 1992; Weinberg, 1992). However, hardly any attention has been devoted to the issue of the *type of assessment system*

envisaged, or the role of measurement in the new education system in South Africa.

Bam and Rice (1987) note:

As far as we know, the specifics of testing have not been addressed, but whatever form they take it will be within the spirit of non-elitism and consultation that lie in the heart of People's Education. How this is going to be achieved remains in the realm of speculation. Given that some testing is being contemplated it will, as any good test must, discriminate one learner from another (p. 5).

1.4.1 Assessment and Education in South Africa

Assessment practices in South African society have historically been, and still are, intimately linked to the social order of the day, and were primarily developed from within a tradition of psychology (Whittaker, 1990). The analysis presented in this section is based on these two themes. First, it is argued that tests and testing instruments are a cultural product, developed to serve specific goals in society. This process can by no means be regarded as neutral, especially in situations where the political, social and economic relations between different groups are characterized by dominance and control of the minority over the majority as well as language and cultural differences. Second, the link between the use of assessment practices and the promotion and maintenance of the status quo is demonstrated.

The rise of the testing movement in South Africa is strongly related to the development of tests in Europe and the United States, and has a similar history. From the very beginning (intelligence) testing in South Africa has been used to produce theories of intellectual differences between races (Apple, 1989; Bulhan, 1981; Swartz, 1992; Whittaker, 1990), a classical example is that of Cyril Burt who used fabricated data to support his belief that Black people had inherited inferior

brains (Whittaker, 1990). Apple (1989) notes that Fick (1929), a South African psychologist trained at Harvard University, reported that the mean intelligence of tested Black children was so low that they almost coincided with the scores of mentally deficient children. With regard to his research on the "Educability of the South African Native" Fick (1929, cited in Whittaker, 1990) concluded that

... the inferiority of the Native (African) in educability as shown by the measurement of their actual achievement in education, limits considerably the proportion of Natives who can benefit by education of the ordinary type beyond the rudimentary level (p. 56).

In 1930, MacCrone, a South African psychologist, (cited in Apple, 1989) noted:

that common perception held by White South Africans, including scientists, was that Blacks had low intelligence, limited knowledge, acceptance of poor standards of living and occupation, criminal tendencies, childishness, and ridiculous behavior. In short, opposite of all those qualities that Whites possess (p. 551).

In an article in the South African Journal of Science (1921, cited in Apple, 1989), J. Duerden, the President of the South African Association for the Advancement of Science, argued for the study of the many different races and nations settled within the borders of South Africa *at such diverse stages of social evolution* (my emphasis).

Duerden further notes that

By virtue of his higher intelligence, not only because of absence of color, I see the White man leading in South Africa: he will constitute an aristocracy of ability, benevolent to the races less endowed (cited in Apple, 1989, p. 550).

The crucial point is that to a large extent, such perceptions of Blacks still prevail quite strongly, especially among those White South Africans who are

responsible for development and implementation of the relevant ideology and philosophy that governs these perceptions and practices. Many of these individuals still hold (often appointed for life) key positions in all state funded (and some private) institutions³, for example universities and educational departments, where most of the assessment practices are conducted.

In addition, the structure of the education system is such that the continuation of these racist practices can be maintained. The education system is still very much segregated according to racial and ethnic lines (see Figure 1.1), where the primary role (of formal education) is to serve as a mechanism of social control and provide a source of cheap [Black] labor (Mathonsi, 1988). This has taken the form of an elaborate system of tests and examinations by which control and entry into the economy is regulated (Swartz, 1992; King & v d Berg, 1992). To a great extent, tests have been intentionally misused to deprive Blacks access to resources and opportunity, and their intellectual development has been stifled in a conscious and systematic manner to meet the needs of the White minority. These needs have mainly tended to be in the form of a cheap source of labor. It is thus no surprise that the emphasis in education is geared towards rote learning, and thus very exam orientated, for both Black and White South Africans. In fact, the development of critical thought and active student participation in the learning-teaching process is

³In the past 5 years concerted efforts have been made to get control of educational institutions into the hands of local Black communities. These efforts have succeeded to a large extent as many 'ethnic universities' have now declared themselves non-racial institutions. However, schools and teacher training colleges are still entrenched in the old apartheid order and continue to operate as such.

actively discouraged. Rather, students are viewed as mere passive receivers of information (Kallaway, 1984).

Due mainly to this history, a strong argument for the total abolition of all forms for assessment practices in South Africa, especially for classification and selection, is by no means unreasonable. However, current conditions, specifically the severe lack of resources, in South Africa dictate the use of some form of selection. Also, relevant and appropriate assessment practices would most certainly play some role in a post-apartheid, democratic, non-racial, non-sexist, non-exploitative society. A case in point is Cuba, a country that has gone a long way towards creating the type of society that many South Africans envisage. Assessment practices are utilized for the ultimate benefit of the society as a whole (that is to increase efficiency in industry, or to promote fairness in the allocation of state resources, etc.), and include all affected members of the community in most decision making, for example, university entrance (Kanjee, 1993b).

Addressing the issue of certification, Swartz (1992) noted that the crucial problem relating to assessment practices in South Africa is:

to evolve a system of certification which will reflect, as far as possible, the interest of the majority of the people - especially the most dispossessed 'racial' groups and social classes. That such a system will favor specific interest, and indeed therefore give effect to a particular 'bias' is self-evident. The issue at stake is to determine the structure of the new 'bias' in a democratic or rather, 'democratizing' social order and in a manner which serves to reduce inequalities (p. 140).

Swartz (1992) also notes that South Africans should apply this "bias towards a democratized social order" towards other forms of assessment as well.

The point is that the primary goals and objectives of assessment, whether for selection or diagnosis, should be for the elimination of existing social inequalities. In this context, assessment practitioners need to be accountable and committed to making these goals and objectives concrete, such that the perceptions and practice of equality and fairness are propagated and maintained for all groups concerned. These goals and objectives must be made explicit to every individual assessed so that he or she can feel confident that they will not be discriminated against or unfairly treated in any way possible, either by responding to ('bias') assessment devices, or by the manner in which results are interpreted and utilized. This is especially important considering the history of assessment practices in general, and specifically, in South Africa.

It must however be emphasized that long term, fundamental changes cannot occur through the use of relevant assessment practices alone. These changes must occur with corresponding changes in the social, economic and political structure as well, changes that complement and reinforce each other for the improvement of the society as a whole. For example, the language question in South Africa is a widely debated, and highly sensitive issue. As long as the "national language question" is not appropriately addressed and resolved, assessment practitioners will continue the use of past practices that required that tests be only made available in the two "official" languages, even though this might not be appropriate of the overwhelming majority of the people. This question of language in educational assessment is addressed in the next section.

1.4.2 Educational Assessment and Translation

A crucial question currently facing South African educators is that of language. Desai (1992) notes that the language policy formulated in South Africa has to assist in the goal of creating a democratic, non-racial, non-sexist and non-exploitative society. She further states that the language policy should not perpetuate the type of division such that

it is no longer claimed (at least not openly) that certain 'races' are more fit to rule than others. Today it is certain ethnic groups, cultures and languages which are claimed to be fitter to rule than others (Skunab-Kangas, cited in Desai, 1992, p.118).

However, this is much easier said than done. Currently, there exists at least eleven different languages, many of which are used by a significant number of South Africans (Bam & Rice, 1987; Omotoso, 1994; Appendix A). Given the fact that: (1) the state had enforced English and Afrikaans⁴ as official languages and thus most people have been forced to learn these languages against their free will, (2) almost 45% of South Africans, mainly Blacks, are illiterate (see Figure 1.2), in their home (mother) languages as well as the official languages (Desai, 1992), (3) there are over ten other Asian and European languages used extensively by the different ethnic groups (Bam & Rice, 1987; C.S.S., 1992), and (4) there are many different varieties of South African creole that incorporate parts of many aspects of the different languages (Desai, 1992), the extent of the problem of determining a

⁴A local South African language derived mainly from Dutch.

national language is quite complex. Appendix A contains a list of the percentage of people using different languages in South Africa.

Some of the possible solutions listed by Desai (1992) include the implementation of a multi-lingual policy with English as the primary language of communication, the standardization of similar varieties of languages, and the implementation of a single official language (most likely English) with other languages given equal status on a regional basis. Whatever the final outcome of this debate, it is evident that more than two languages will be used in the future South African society.

The implication for educational assessment is that appropriate strategies need to be made available to ensure that when assessed, all South Africans can fully and equally participate in this process. In this respect, translation of assessment devices could provide a unique solution, especially as the proper use of translation practices in assessment can enhance the equal and fair treatment of all the different ethnic and language groups in South Africa. However, there are other more fundamental problems that need to be addressed before an acceptable system of evaluation that serves to promote the ideals of equality and fairness to all South Africans can be established.

1.4.3 Possible Problems and Obstacles

One of the fundamental problems regarding assessment in South Africa is the lack of suitably trained and qualified assessment practitioners and specialists, especially from and for Black communities (Malaka, 1992; Mathonsi, 1988). The

consequence of this is that most assessment instruments currently used in South Africa have been (or still are), adapted from other countries, mainly the U. S. and Britain (Prinsloo, 1984). Primarily, these adaptations and translations are available in English or Afrikaans only (HSRC, 1993). In addition, almost all assessment devices for Blacks have been developed by Whites, or those few that either involved participation of Blacks or have been developed by Blacks, have had to be approved by the [White's only] registration body (HSRC, 1993). Thus at present, very few instruments exist which specifically address issues relevant to Black South Africans. Exacerbating this situation is the fact that even fewer instruments exist that are able to compare different 'racial/ethnic' groups such that the practice of equality and fairness to all groups is maintained, for example, testing instruments which allow students to use their first (home) languages instead of their second or third language.

Another major problem specific to South Africa is the lack of qualified and trained translators. While many South Africans speak more than two languages (English, Afrikaans and the mother language), sometimes even four languages, there are very few who are trained or qualified as translators. This is especially true of the indigenous African languages - as noted by the lack of any academic training program focusing on African languages. Even the 'ethnic universities' use English as a medium of instruction. The implication of this on assessment practices is crucial as use of poorly trained translators in developing instruments can affect score reliability and validity (Hambleton & Kanjee, in press).

The lack of content and curriculum specialists, especially from Black communities, is another issue that poses a major problem, as these individuals, as

qualified and experienced translators and item writers, play a significant role in the development of any assessment device. Fortunately, the number of Black South Africans focussing on this area has steadily increased in the past few years as it was, and still is, seen as a crucial component for the new education system. However, there is still a need for greater participation in this area (Jansen, 1991).

The issue of funding for education and education assessment in particular could pose another major obstacle. With the emphasis of building a new society, current problems dictate that priority be given to the provision of basic needs first. In this respect, the issue of housing, employment and education need to be immediately addressed. However, even within the educational arena, the easy access to free education for all South Africans (at least at the primary and secondary levels), the provision and upgrading of schools, especially for Blacks, the introduction of relevant curricula, upgrading and training of more qualified teachers, etc., would surely take precedence and priority over the improvement and introduction of relevant assessment methodology. Thus it is not an unreasonable prediction that assessment specialists would almost certainly be expected to work in environments with restricted access to funding. The implication of this is clear in that it severely restricts: (1) new research, (2) the introduction of new and innovative methods, and (3) the provision of the best possible assessment alternatives for all concerned. Assessment practitioners thus need to devise alternative and innovative ways of cutting costs, yet keeping abreast of new developments. Adapting existing instruments (that are relevant, reliable and valid) as opposed to developing new instruments, provides a unique way of achieving this.

While equal access to relevant education for all South Africans constitutes one of the basic premise of the new society, currently almost all educational institutions are still very much segregated along ethnic and racial lines (see Appendix B). Even when all educational institutions are eventually open to all South Africans, the effect of differential schooling and access to vastly different facilities needs to be taken into account. Of particular interests to assessment practitioners is the effect of vastly different curricula, teacher qualifications, educational facilities (for example laboratory equipment), class sizes, as well as exposure to and experience with tests and testing practices (for example essay versus multiple choice, written versus oral, being assessed in second or third languages as opposed to first language, physical conditions under which assessment occurs, etc.). All these factors need to be taken into account whenever any individual is assessed, whether for selection or for diagnostic purposes.

1.5 Purposes of the Study

The problems and issues noted above are few of the many facing South African (and international) assessment practitioners, and are indeed crucial and serious. Their solutions are by no means impossible to achieve, and could provide a significant contribution towards the attainment of more relevant and valid information regarding cross-national/cultural comparisons, especially in the field of education. It is acknowledged that the move towards (positive) progress is a gradual process, and requires the concerted effort and dedication of many people. Given this, the specific purposes of this study were to:

1. provide a comprehensive review of the test translation literature, and
2. investigate the impact of several methodological variables that impact on cross-cultural studies, as it pertains to DIF detection in multiple language group comparisons.

The hope is that by addressing these two purposes, the study would: (1) contribute to the growing number of studies which provide knowledge and information on cross-cultural comparisons, especially as it relates to educational and pedagogical issues, (2) sensitize assessment practitioners to the history of assessment practices and its negative role in the oppression of "minorities", as well as the positive role that these practices can play in society, and (3) caution practitioners that no matter how sound and valid available techniques are, the interpretation and manner in which information is utilized and reported is just as crucial. In this respect, assessment practitioners have a responsibility to ensure that the development of new techniques and procedures, and interpretation and reporting of information is conducted in a manner that not only is sensitive to the various histories of the different cultural/national groups, but that would promote the ideology of mutual benefit and cooperation towards eradicating social inequalities between different language or cultural groups.

1.6 Organization of the Thesis

Chapter 2 contains a review of the test translation literature. Hence, some of the reasons for translating tests, the various judgmental and statistical methods used to assess equivalence or identify DIF, the specific statistical techniques used, as well

as the specific questions that this study addresses will be introduced. A study to compare the performance of two estimation procedures, item response theory and logistic regression procedures, for identifying DIF in multiple groups is described in Chapter 3. Chapter 4 provides the methodological details and results for a computer simulation study to investigate the effect of two variables, sample size and ability distribution on the performance of the two estimation procedures. Chapter 5 includes a summary of the findings and recommendations from this study.

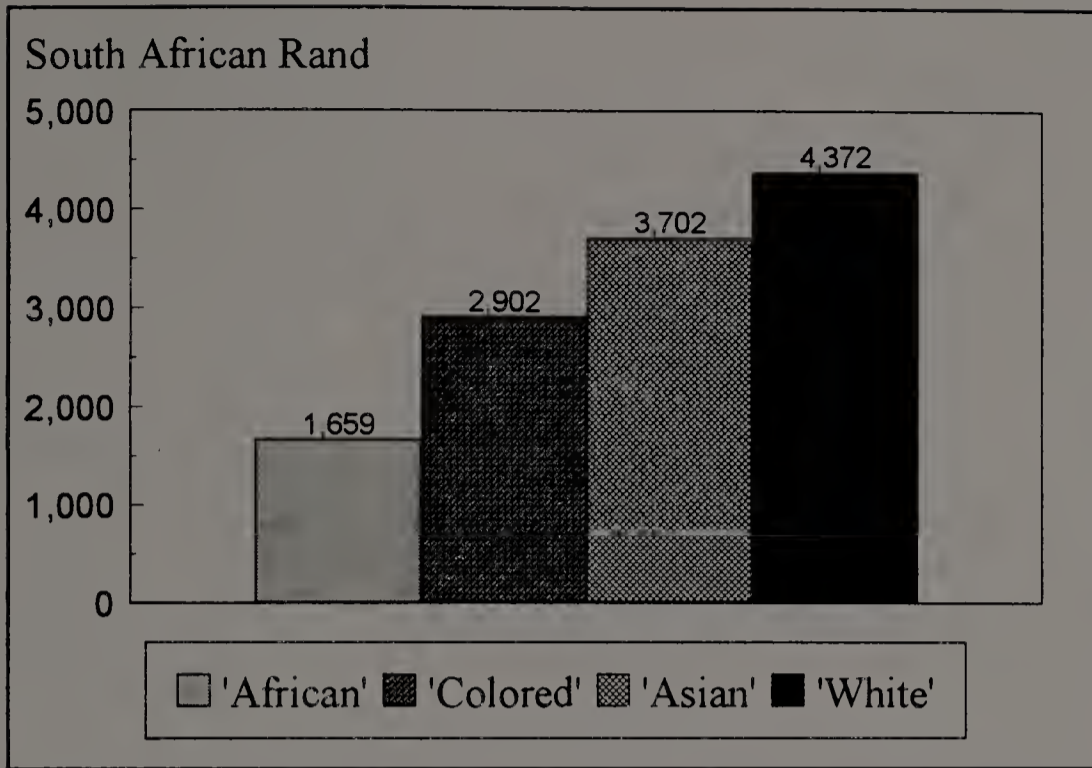


Figure 1.1

Education funding in South Africa for 1992-1993

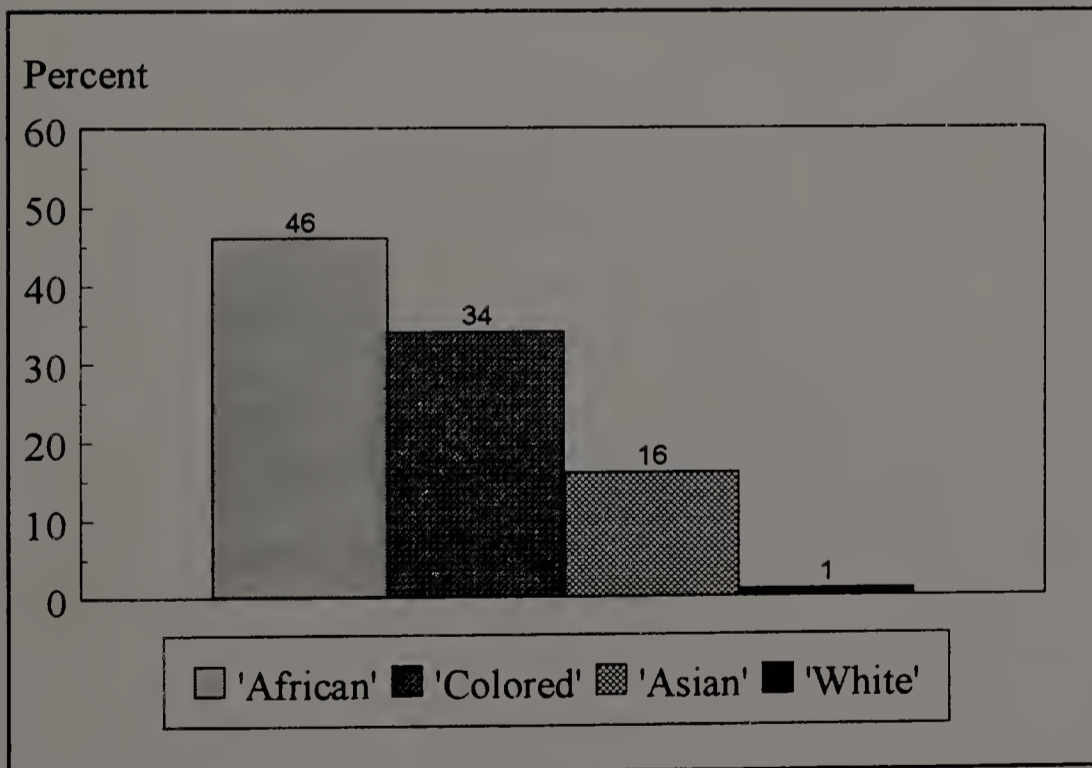


Figure 1.2

Illiteracy rates in South Africa for 1993

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Introduction

The different designs and methods used to assess equivalence of translated instruments are considered in this chapter. Some of the major reasons for translating and/or adapting assessment instruments are first noted, and then a definition of equivalence is presented. Popular judgmental methods and statistical methods, for which a distinction is made between statistical designs and statistical techniques, are presented. The statistical techniques noted are discussed in the context of identifying differential item functioning (DIF) and focus specifically on item response theory (IRT) and logistic regression (LR) procedures.

The methods and designs used for establishing equivalence between the original and translated versions of an instrument can be viewed as an extension of the methods used for identifying DIF (Hambleton & Bollwark, 1991). In DIF studies the focus is on the scores derived from items on a *single* testing instrument in two (or more) groups of interest, while in translation studies that assess equivalence between two or more instruments, the focus is on the scores derived from items from *different* testing instruments (that is, the original as well as the translated/adapted instrument administered to different groups). In this study however, identifying DIF and assessing equivalence can be used interchangeably as in both cases, the focus is on comparing items in a test instrument when multiple language versions exist.

It must, however, be noted that the DIF detection procedures applied will inevitably flag items for other reasons as well, for example, unfamiliar formats and inappropriate content. The point is that DIF studies are broader in their approach to assessing the validity of test instruments in multiple groups than studies designed solely to look at problems created by the translation of test instruments.

2.2 Reasons for Translating Instruments

The reasons for translating and/or adapting assessment instruments include:

1. To enhance fairness in assessment by allowing persons to be assessed in the language of their choice. For example, in a recent study to measure verbal and non-verbal abilities in South Africa, students were given the choice to take the instrument in any one of five languages (A.R. van den Berg, personal communication, June, 1993). Thus possible bias associated with assessing students in their second or third best language is removed and the validity of results increased.
2. To facilitate comparative studies across national, ethnic and cultural groups, both at an international and national level. This is especially relevant in recent years with the growing contact and cooperation of different nations in economic, educational and cultural spheres, and has resulted in an increased need for many nationalities and groups to know and learn more from and about each other. For example, over 60 countries will participate in the Third International Mathematics and Science Study (TIMSS) being conducted in 1995 and 1999 (Hambleton & Kanjee, in press).

3. To reduce costs and save time in developing new instruments. It is often cheaper and easier to translate and adapt an existing instrument into a second language than to develop a new instrument. This is especially true in situations where there is a lack of resources and technical expertise.

Brislin (1986) also notes that instruments are translated to enable the use of existing data. That is, to use existing norms, interpretations and other information which can be compared with the newly acquired information. An additional reason is the sense of security that established and respected instruments provide.

However, Brislin cautions against using an assessment instrument merely because it is well established since reliability and validity of the instrument in the target culture are not guaranteed.

While the use of translations in testing has been common from the time of Binet's first test in the early 1920's (Popham, 1990), it must be recognized that the field of cross-cultural and cross-national comparisons in education, is still relatively new and in its infant phase of development. Currently, the main concerns revolve around methods and designs for translating instruments, focussing on establishing the equivalence of scores (Drasgow & Hulin, 1986; Ellis, 1991; Hambleton, 1993; Poortinga, 1983; Van de Vijver & Poortinga, 1991), ways of interpreting and using data (Hambleton & Kanjee, in press; Poortinga & Malpass, 1986), and defining standards and guidelines for translating assessment instruments (Hambleton, 1993).

In this study, the primary focus is on the methods and designs used for assessing equivalence in translated assessment instruments. Specifically, this study addressed the identification of DIF when multiple groups from different language

and/or cultural backgrounds are compared. However, some information regarding the definition of equivalence and DIF would prove helpful in understanding these different designs and methods of assessing DIF discussed in later sections.

2.3 Equivalence and DIF in Cross-Cultural/Language Comparisons

The attainment of equivalent measures is perhaps the most central issue in cross-cultural/national comparative research (Poortinga, 1983). If the basis of comparison is not equivalent across different groups then valid comparisons across these groups cannot be made. Certainly, observed scores from the groups taking different instruments are on different scales and are thus not directly comparable (Drasgow & Kanfer, 1985; Lonner, 1990). For any comparison between different language/cultural groups to be valid, all measurement instruments used must be demonstrated to be equivalent. Hulin (1987) provides the following definition of equivalence:

If individuals with the same amounts of the trait being estimated have different probabilities of making a specified response to the item when responding to different language versions of the items or scales, the items are said to be biased or nonequivalent (p. 138).

That is, individuals with the same standing on a construct, say math ability, but belonging to different groups, say Brazilians and Nigerians, should have the same expected observed score on the instrument measuring that construct. However, even if scores are comparable, it cannot be assumed that the instrument is free of any DIF.

Defined within the framework of differential item functioning (DIF), two or more versions of an item prepared in different languages are assumed to be

equivalent when members of each group of the same ability have the same probability of success on the item (Hambleton, Swaminathan, & Rogers, 1991). It must be noted that there is no requirement for equal distributions on the construct being measured across the different groups (Drasgow & Kanfer, 1985). Thus, within any cross-cultural/national comparison, it is possible for some groups to display higher or lower scores than other groups. In this context, it is important to ensure that any score differences manifested between these groups are not due to the failure of the measurement instrument to provide equivalent scores, and thus the identification and elimination of any DIF is crucial.

It is important to note that the identification and elimination of DIF from any instrument increases the reliability and validity of scores. Thus, results from two or more groups from which DIF items have been removed are more likely to be comparable and thus equivalent. One of the ways of increasing this likelihood is to ensure that appropriate methods and designs are applied whenever two or more groups from different language and/or cultural backgrounds are compared. In the next section, the judgmental and statistical methods of assessing equivalence are presented.

2.4 Judgmental Designs for Assessing Equivalence

Judgmental designs of establishing translation equivalence are based on a decision by an individual or a group of individuals on the degree to which items are equivalent in the source and target languages (Hambleton & Bollwark, 1991). The

methods discussed are distinguished by the use of (1) forward-translation, and (2) back-translation designs.

2.4.1 Forward-Translation Designs

In this design, the source version of an instrument is first translated into the target language by a single translator or a group of translators (Hambleton, 1993). In one variation of this design, one or more samples of target examinees answer the target version of the instrument and are then questioned by judges about the meaning of their responses. Judges decide if the responses given reflect a reasonable representation of the item in terms of cultural and linguistic understanding. If a high percentage of examinees present a reasonable representation of an item (in the target language), the item is then regarded as being equivalent to the source language. The main judgement here is whether the target language examinees perceive the meaning of each item on an instrument in the same way as the source language examinees (Hambleton, 1993). A common variation of involves judges studying the source and target language versions of a test to assess equivalence. Changes can be made in a translated instrument based upon information provided from the judgmental review.

The advantage of this version of the forward-translation design is that valuable information about the functioning of any item is provided directly by the examinees, information that is otherwise unavailable when examinees only respond to questions on paper. However, the disadvantage is that there are many factors (personal, cultural, linguistic) during the interaction between examinees and judges

that can quite easily interfere with the results. For example, judges can easily misinterpret, misunderstand and/or misrepresent responses of target examinees. Another disadvantage is that this method is that it is labor intensive and time consuming compared to other judgmental methods (Hambleton, 1993). The third problem is that if the instrument used by source language monolinguals is not valid or the meaning of responses from examinees are not fully understood, comparing the results to target language monolinguals is meaningless. That is, one has to be certain of the meaning of responses from source language monolinguals before judging responses from target language monolinguals, as the former (subjective) interpretations of the responses of examinees consists of the basis by which the latter responses are judged.

In the more common variation of this design, instead of having target group examinees answer the translated version of the instrument, a single (or preferably a group of) different translator(s) compare the source and target versions of the instrument to determine whether the two versions are equivalent. Hambleton and Bollwark (1991) note that these comparisons can be made on the basis of having translators simply look the items over, check the characteristics of items against a checklist of item characteristics that may introduce non-equivalence, or by having them attempt to answer both versions of the item before comparing them for errors. Three problems of this variation are: (1) it is often difficult to find bilingual judges who are equally familiar with the source and target languages and/or culture, (2) bilingual judges may inadvertently use insightful guesses to infer equivalence of meaning, and (3) bilingual judges may not think about the item in the same way as

the respective source and target language monolinguals and thus the results may not be generalizable.

2.4.2 Back-Translation Designs

In back-translation designs, the original instrument is first translated into the target language by a set of translators, and then translated back into the original language by a different set of translators (Brislin, 1986). Equivalence is usually assessed by having source language judges check for errors between the original and back-translated versions of the instrument. The main advantage of this design is that researchers who are not familiar with the target language can examine both versions of the source language to gain some insight into the quality of the translation (Brislin, 1976). Also, this design can easily be adapted such that a monolingual researcher (assessment or subject specialist) can evaluate (and thus improve) the quality of the translation after it has been translated into the target language, but before it is back-translated into the source language.

The main disadvantage of this design is that the evaluation of instrument equivalence is carried out in the source language only. It is quite possible that the findings in the source language version do not generalize to the target language version of the instrument. This might happen if the translators use a shared set of translation rules that insures that the back-translated instrument is similar to the original instrument (Hambleton, 1993). Another disadvantage is that the assumption that errors made during the original translation will not be made again during the back-translation is not always applicable (Hambleton & Bollwark, 1991). Often

skilled and experienced translators use 'insight' to ensure that items translated are equivalent, even though this may not be true. This, however, can be controlled by either using a group of bilingual translators or a combination of bilinguals and monolinguals to perform multiple translations to and from the target and source languages (Bracken & Barona, 1991; Brislin, 1986). For example: (1) Brislin (1986) suggested the use of monolinguals to check the translated version of the instrument and make necessary changes before it is back-translated and compared to the original version; (2) once the two versions of an instrument are as close as possible, Bracken and Barona (1991) suggested the use of a bilingual committee of judges to compare the original (or back-translated) and the translated version of the instrument to ensure that the translation is appropriate for examinees.

The judgmental methods include the use of (1) forward-translation, and (2) back-translation designs. Both the designs can certainly provide researchers with valuable information about the equivalence of measuring instruments. However, the sole use of judgmental designs for assessing equivalence does not provide adequate evidence of equivalence because no examinees ever see the two versions of the instrument. Since examinees are often operating at a different cognitive level than translators (and under test-taking conditions), it is highly possible that the translation found to be acceptable by translators (judges) may not actually be so in practice. Hambleton (1993) notes that most of the available evidence suggests that judges are not very successful at predicting items on an instrument that function differentially in two or more groups. To this end, the suggested practice is that judgmental methods should be supplemented with appropriate statistical methods as well

(Bracken & Barona, 1991; Hambleton, 1993; Prieto, 1992). In the next section, these statistical methods are discussed.

2.5 Statistical Methods of Assessing Equivalence

The statistical methods employed to identify DIF between two (or more) assessment instruments in different languages are characterized by the (1) statistical design, and (2) statistical technique used. The statistical design used is dependent on the characteristics of participants (that is, monolingual or bilingual), and on the version of the translated instrument (that is, original, translated or back-translated) (Hambleton & Bollwark, 1991), while the statistical technique(s) selected are dependent on whether a common scale is assumed and whether conditional or unconditional procedures are applied (Van de Vijver & Poortinga, 1991). These factors determine the specific analytical techniques (factor analysis, item response theory, logistic regression, etc) best suited to identify DIF, and thus a thorough and complete understanding of the current applicable statistical techniques is vital. In the next section, a brief explanation of some of the applicable statistical designs and techniques currently used is presented.

2.5.1 Statistical Designs

The three statistical designs discussed in this section are based on whether examinees used to assess item equivalence are: (1) bilingual, (2) both source and target language monolinguals, or (3) only source language monolinguals.

2.5.1.1 Bilingual Examinees

In this design, both the source and target versions of the instruments are administered to bilingual examinees, and the two sets of scores are then compared. Care is taken to ensure that the order of instrument presentation is counter-balanced and that the time between administrations is short enough that ability scores are not likely to change. The advantage of this design is that since the same examinees take both versions of the instrument, differences in the abilities of examinees that can confound the evaluation of translation equivalence will be controlled (Hambleton & Bollwark, 1991). The disadvantage of this design is that due to time constraints, examinees might not be able to take both versions of the instruments. A variation of this design that overcomes this problem of time is to split the bilingual sample and randomly assign examinees to only one version of the instrument. Now, the item and instrument performance of the randomly equivalent groups can be compared.

However, the problem of differences between the examinees with respect to their 'level' of bilingualism and/or 'level' of biculturalism could still violate the assumption of equal abilities between examinees (Hambleton, 1993). Another more serious problem is that the results obtained from bilingual examinees may not be generalizable to the respective source language monolinguals (Hulin, 1987). Also, with regard to cross-national studies, the use of this design is not a feasible option as it is very difficult to find individuals who are equally familiar with the cultures and languages of the nationalities being compared. Language dominance tests are available for the most common languages, and they could be helpful, but concerns about their validity exist.

2.5.1.2 Source and Target Language Monolinguals

In this design, source language monolinguals take the source version and target language monolinguals take the target version of an instrument (Brislin, 1986; Candell & Hulin, 1986; Ellis, 1989; 1991; Hulin & Mayer, 1986). The source version can either be the original or back-translated version of the instrument (Brislin, 1986). The two sets of scores are then compared to determine the equivalence between the two versions of the instrument. The main advantage of this design is that since both source and target language monolinguals take the versions of the instrument in their respective languages, the results are more generalizable to their respective populations. A major problem is that since two different samples of examinees are compared, the resulting scores may be confounded by real ability differences in the groups compared (Hambleton, 1993).

However, alternative steps can be taken to minimize this problem (Bollwark, 1991). First, examinees selected for the groups should be matched as closely as possible on the ability/abilities of interest. Matching should be based on criteria that are relevant to the purpose of assessment. For example, scores from instruments that assess correlated tasks/abilities could be used. If such information is unavailable, examinee samples should be chosen using the most available information about the ability level of each sample, for example, years and type of schooling and/or demographic data may be used. Second, conditional statistical techniques that take into account the ability of examinees when comparing test scores on an instrument can also be used to control for ability differences in the

source and target examinee samples, for example, methods based on item response theory and/or logistic regression.

Last, factor analysis or any other statistical technique where no common scale is assumed are often used in conjunction with this design. For example, in factor analysis, scores of the two groups are separately analyzed to determine the similarity of the factor structures across the two groups. However, the disadvantage is that since factor analysis is based on classical item statistics, the results are sample dependent (Hambleton & Bollwark, 1991). Still, researchers must check that the ordering of item difficulties is the same in the two versions of the instrument.

2.5.1.3 Source Language Monolinguals

In this design, equivalence of the instrument is based on the scores of source language monolinguals who take both the original and the back-translated versions of the instrument. The advantage is that the same sample of examinees is used and thus scores are not confounded by examinee differences. A major problem, however, is that no data on the performance of target language individuals, nor the translated version of the instrument is collected. Thus information about possible problems concerning the target group version is not available, making the validity of this design very limited.

2.5.2 Statistical Techniques to Detect DIF

Statistical techniques based on IRT are considered by many researchers to provide a more theoretically sound approach for the study of DIF (Linn & Harnisch,

1981; Shepard, Camilli & Williams, 1985). In addition, Scheuneman and Bleistein (1989) note that the three-parameter IRT model is preferred over the one- and two-parameter models. The major advantage of IRT methods with regard to the detection of DIF is the property of population invariance. However, a major disadvantage, especially for the two- and three-parameter models, is that relatively large sample sizes are required for estimating parameters which are sometimes difficult to attain in practice (Zieky, 1993).

A review of the psychometric literature indicates that many alternatives to the IRT based procedures have been proposed. A discussion of all these techniques used to detect DIF or to assess item equivalence is beyond the scope of this study. Some of the more popular techniques that are currently used include logistic regression (Bennett, Rock & Kaplan, 1987; Swaminathan & Rogers, 1991), the Mantel-Haenszel procedure (Clauser, 1993; Hambleton & Rogers, 1989; Holland & Thayer, 1988; Schmidt, Holland & Dorans, 1993), the standardization procedure (Dorans, 1989; Dorans & Holland, 1993), and factor analytic procedures (Knoll & Berger, 1991; Mayberry, 1984; Royce, 1988; Triandis, 1976).

It must be noted, however, that the Mantel-Haenszel (Holland & Thayer, 1988) and the logistic regression (Swaminathan & Rogers, 1990) procedures are perhaps the best known of the many non-IRT based alternatives proposed to detect DIF (Rogers & Swaminathan, 1994). Compared to IRT procedures, both these procedures are easier to use and understand, are readily available, are applicable to relatively small sample sizes, and are associated with significance tests to aid in interpreting the DIF statistic (Hambleton & Rogers, 1989).

2.5.2.1 Simultaneous Detection of DIF in Multiple Groups

DIF detection techniques have been applied primarily using pairwise comparisons. However, in some situations it may be necessary to assess DIF in more than two groups, for example cross-cultural or cross-national studies. In this context, pairwise comparisons may be problematic. Multiple pairwise comparisons can prove to be very time consuming and costly. For example, in South Africa where there are 11 official language groups, the use of pairwise comparisons would entail 55 separate comparisons. As a second example, Keeves (1992) noted that a total of 24 countries participated in the Second IEA Science Study. This would require an even greater number of comparisons. The process is further complicated because (1) the items flagged as DIF may differ with each comparison and since all flagged items need to be accounted for, DIF studies in many real life situations would become absurd, and (2) the need to apply the two-stage procedure (Holland & Thayer, 1988) would double the number of comparisons, yet again.

A possible solution would be to assess DIF simultaneously in all the groups, or at least reduce the number of comparisons conducted without excluding or eliminating any group(s) from the analysis. Unlike pairwise comparisons, in simultaneous comparisons, the data from all the different groups in the study are always included in all comparisons that are conducted. For example, when comparing 4 groups, the existence of DIF is determined by comparing group 1 to groups 2, 3 and 4 combined. The advantages of this is that (1) provided the total sample size is large, analysis can be conducted on many different groups even though some of these groups may consist of relatively small sample sizes, (2) the

total number of (multiple) pairwise comparisons are reduced, especially when many groups are compared, and (3) the performance of different individual groups can be directly compared to that of, what Ellis and Kimmel (1992) call the "composite group" (that includes the combined responses of all groups in the study), which serves as a point of reference.

Ideally, only a single estimation should be required to detect any DIF, either for or against any of the groups in the study. In this respect, the use of IRT based procedures seems ideal. Ellis & Kimmel (1992) used IRT procedures to determine 'unique cultural responses patterns' of English, German and French subjects regarding their attitude towards mental health. In their study, Ellis & Kimmel (1992) combined the responses of all subjects into what they called the "omnicultural composite" group, which served as a reference against which the responses of all other groups were compared. Items which differed significantly from the omnicultural composite were indicative of cultural responses that are unique for the group under investigation. With regard to DIF, Ellis & Kimmel (1992) noted that:

in the same way that a DIF item identified a two-group comparison indicates that an item functions differentially for the two groups, a DIF item identified in a comparison between an individual culture and an omnicultural composite indicates that an item functions differentially for the individual cultures compared to the omnicultural composite (p. 178).

If the approach used by Ellis & Kimmel (1992) is adopted for DIF studies involving multiple groups, all items that are flagged for any specific group could simply be regarded as being in favor of or against a specific group. Removing (or revising) these DIF items result in equivalent scores, and thus the performance of groups studied can be directly compared. The point is that the omnicultural

reference group, free of any DIF, represents an underlying construct that is (equally) common to all the groups studied, and thus no single group would have any advantage over any other group, for whatever reason, for example, differences in access to resources, different curricula, etc. However, besides the Ellis & Kimmel (1992) study, data regarding the performance of currently available techniques to simultaneously detect DIF in multiple groups are not generally available. Also, very little information in the literature suggests that any research to this end is being conducted. In this study, the use IRT procedures to detect of DIF in multiple groups is investigated.

Of the current DIF detection methods, only the IRT and logistic regression procedures were selected for study. While the many advantages of the Mantel-Haenszel procedure are recognized, this procedure cannot be used for the simultaneous detection of DIF in more than two groups. The pseudo-IRT procedure proposed (denoted PIRT) by Linn and Harnisch (1981) was selected for study. Shepard, Camilli and Williams (1985) compared the PIRT method to the chi-square and Angoff-delta plot method using real and simulated data. The PIRT method was the method of choice when sample size in the minority group was small (300 or less). The authors noted that the PIRT method was more accurate than the chi-square method and is highly correlated with the widely accepted three parameter ICC method.

The PIRT procedure can easily be extended to detect DIF in three or more groups. The item parameter estimates used to obtain the ability estimates for the respective target groups can be computed based on the sample of examinees which

includes all the groups in the study. The logistic regression procedure (denoted LR) proposed by Swaminathan and Rogers (1990) was also used in this study, as it not only has all the advantages associated with the PIRT procedure, but it is easier to understand, and less time consuming and costly to run than the IRT-based DIF detection procedures. These procedures are discussed in the next section.

2.5.2.2 Pseudo-IRT Procedure

Before the PIRT procedure is discussed an overview of IRT will be presented. Some of the advantages and disadvantages of IRT are listed and its specific application to DIF detection are noted. The PIRT procedure, as proposed and used by Linn and Harnisch (1981) is then discussed. The advantages, disadvantages and some of the factors that affect the PIRT procedure are noted as well.

Item response theory (IRT) models specify the mathematical relationship, known as the item characteristic curve (ICC), between an examinee's latent trait and observed performance on a instrument designed to measure that (latent) trait. Hambleton, Swaminathan & Rogers (1991) note that IRT is based on the assumption that: (a) the probability of a response is a function of only a single trait; that is, the instrument measures one and only one trait, (the assumption of unidimensionality). Equivalently, the examinee's performance on any pair of items are statistically independent; that is, the examinee's performance on one item does not in any way affect performance on another item when conducted on the trait under investigation

(assumption of local independence), and (b) the item response model is adequate in the sense that it fits the test data to which it is applied.

Some of the advantages of the item response theory approach are that it provides a model based approach to assessment (Wainer, 1993a), it takes into account the continuous nature of examinee abilities, and results in invariant item parameters. The property of invariance allows for the direct comparison of the parameters between the groups under investigation (Hills, 1989). However, one of the major disadvantages of IRT models is that the precision of item parameter estimates are influenced by the sample sizes of examinees (Hambleton & Cook, 1983). In the absence of large sample sizes these estimates may have large errors (even with large samples, problems can arise), thus leading to inconclusive results. Another disadvantage is that parameter estimates are not equally accurate in all regions of the ability scale, and thus if groups differ widely in ability, parameters for one group may be more accurately estimated for one group than the other (Hills, 1989). Lastly, IRT procedures based on the two- and three-parameter logistic models are relatively time consuming and costly to use (Hambleton & Swaminathan, 1985). However, recent technological advances and the availability of appropriate software have reduced the cost and time required to run IRT analyses substantially.

In principle, the property of invariance makes the detection of DIF straightforward. That is, when two groups are compared, the resulting set of item parameters should be identical within sampling fluctuation after proper scaling adjustments (Kim, Cohen & Park, 1993). Usually, the ICCs of the reference and target groups are compared (see Figures 2.2 and 2.3). Since the ICCs are

determined by their item parameters, the ICCs for any item that functions equivalently for the two groups should also be the same. That is, the probability of a correct response for persons at a given ability level should be the same for both groups. If an item displays different ICCs for the two groups, the item is functioning differently and should be flagged as DIF. It is within this context that the operational definition of DIF is used in this study. That is, "an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, et al, 1991, p. 110).

Typically, IRT procedures used to identify DIF include the comparison of item parameter values, the area method, as well as the squared differences and sign tests method (see Hambleton & Swaminathan, 1985). One of the requirements for using these procedures is that relatively large sample sizes are required for obtaining stable item parameter estimates (Hambleton & Swaminathan, 1985). However, minority groups typically consist of small samples, and DIF studies are difficult to carry out well (Clauser, 1993; Parshall & Komrey, 1992; Zieky, 1993). To overcome this problem, Linn and Harnisch (1981) developed a procedure that addresses the issue of assessing DIF with small sample sizes within the framework of IRT.

The PIRT procedure was adopted primarily for use with smaller sample sizes in the target (usually minority) group (Ironson, 1982). In this procedure the item parameter estimates (i.e. item discriminating power, a , item difficulty, b , and the lower asymptote, c) along with the ability estimates were first obtained based on the total sample of examinees which includes members of the target group. P_{ij} , the

estimated probability that person j would answer item i correct was obtained using the formula:

$$P_{ij} = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_j - b_i)}} \quad (2.1)$$

where a_i , b_i , c_i , and θ_j are all estimates.

Next, members of the target group were divided into score categories on the basis of their estimated θ 's. In their study, Linn and Harnisch (1981) used 5 score categories, though more groups are likely to provide a better basis for detecting DIF, target group sample size permitting. The ability groups need to be large enough in size to lead to stable results. In this study, 10 score categories were used (Yen, 1981).

Last, a standardized difference score for examinees in ability group q was computed. In this study, the formula:

$$Z_{iq} = \frac{U_{iq} - P_{iq}}{\sqrt{\frac{P_{iq}(1 - P_{iq})}{N_q}}} \quad (2.2)$$

was used, where i denotes the item, q denotes the ability category, U_{iq} is the observed proportion of correct responses on item i in the q th ability category, N_q the number of persons in ability category q , and P_{iq} is the expected proportion of correct responses obtained using the item response model. Summing across intervals, the overall index of DIF, Z_i is computed as:

$$Z_i = \frac{\sum_q N_q Z_{iq}}{\sum_q N_q} \quad (2.3)$$

The Q_1 chi-square statistic (Yen, 1981) was used to test for significance of fit. The Q_1 statistic for item i is given by:

$$\begin{aligned} Q_{1i} &= \sum_{q=1}^m \frac{N_q (U_{iq} - P_{iq})^2}{P_{iq} (1 - P_{iq})} \\ &= \sum_{q=1}^m Z_{iq}^2 \end{aligned} \quad (2.4)$$

where examinees are divided into m ability categories on the basis of their ability estimates. The statistic for Q_1 is distributed as a chi-square with degrees of freedom equal to $m - k$, where k denotes the number of parameters in the IRT model (Hambleton, Swaminathan & Rogers, 1991). If Z_{iq}^2 exceeds the critical value (obtained from the chi-square table), the item is flagged as DIF.

The use of the PIRT procedure has many potential advantages. First, it enable the use of IRT based techniques to "identify items which could be biased for members of a particular group (target group) when only modest sample sizes are available" (Linn & Harnisch, 1981, p. 115). What the authors fail to specify is just how small the sample can be before rendering this approach invalid. Also, it must be noted that if the total sample size is small (that is the sample size when all the examinees are combined into a single group), the use of IRT is highly questionable (Hambleton & Swaminathan, 1985). Second, the computation of differences (Z_i) in

particular regions in the θ scale may have special appeal in testing situations where there is a special interest or need for further investigation into particular regions, for example, in minimum competency testing. Third, the use of specific score category group differences as well as overall group differences allows one to detect items which show small overall differences yet have large positive and negative differences that tend to cancel out over the range of θ 's. Fourth, situations when simple comparisons using ICCs sometimes suggest the existence of DIF due mainly to the fact that there are relatively few observations for one of the groups being compared are easily avoided. This is because the indices are weighted by the distribution of estimated θ 's in the target groups. Last, the standardized difference score can readily be used to develop a significance test.

Some potential disadvantages were also noted by the authors. First, the estimates of the item parameters are contaminated when the target groups being investigated are included in the total estimation sample. Linn and Harnisch (1981) noted that this contamination will depend on the size of the target group. When DIF does exist, this would tend to reduce the magnitude of the DIF indices. A possible alternative is to use only non-target group examinees to estimate the item parameters and then treat these item parameter estimates as fixed to obtain θ estimates for the target group members. These alternative estimation procedures are discussed in greater detail in the next chapter.

Second, the values of the indices calculated in each ability score group depend on: (1) the number of ability score groups used, and (2) how members of the target groups are assigned into each of these groups. While Linn and Harnisch

(1981) used 5 ability groups, they did not specify how target group members were assigned into each ability group. Possible approaches that could be used to define these ability groups are the equal N, equal θ , and equal probability interval (Hosmer & Lemeshow, 1989).

For all three of these approaches for forming ability groups, the number of score categories (sc) must be defined first, for example, $sc = 10$. In the equal N interval approach, the number of examinees assigned to each score category are equal. For example, if 1000 examinees are in the target group, they are divided into 10 categories of 100 each, based on a ranking of examinee using ability scores. That is, 100 examinees with the lowest ability scores form the first ability group, those examinees with the next lowest 100 ability scores forms the second ability group, and so on. In the equal theta interval approach, the range of estimated theta values are first divided by the number of score categories to define the cutpoints of the theta intervals. Examinees are then assigned into each of the specific intervals on the basis of their estimated θ score. For example, if $sc=10$ and the θ scores range from -1 to 3, the interval width is 0.4. Thus all examinees with theta scores between -1.0 and -.60 are assigned to the first score category, examinees with θ scores between -.60 and -.20 are assigned to the second category, and so on.

In the equal probability interval approach, examinees are assigned to score categories based on the percentiles of the estimated probabilities. For example, assume that 10 score categories are used. For each item, all examinees who have a probability of .1 (or lower) of responding correctly to that item are assigned to the first score category. Examinees with a probability between .2 and .3 are assigned to

the next score category, and so on. Thus, for each item, while the number of score categories remains the same, the number of examinees in each category can differ depending on the relative difficulty of the item.

Last, the authors note that because of sample dependence, these indices cannot be compared from one target group to another across testing situations. That is, the residuals computed are only relevant for the sample of examinees on which they were based. If this sample were to be changed, then calculations need to be redone.

Basically, the pseudo-IRT method uses all available information when faced with small sample groups to (1) estimate the item parameters from the total sample of examinees, and (2) compare the target group residuals to identify DIF. What in fact they are doing is comparing the data fit for the groups under investigation, that is the actual item performance is compared to that predicted from the best fitting IRT model using the total examinee sample. If the fit is poor, DIF is said to exist and relevant items are identified for further investigation. This approach certainly provides an alternative to using IRT in assessing equivalence when sample sizes of one or more groups are relatively small, for example, in cross-cultural comparisons. However, it appears that some of the disadvantages noted could render this method impractical for use. Some of these disadvantages (to some extent) can be overcome by using the logistic regression (LR) procedure. In the next section, the LR procedure is discussed.

2.5.2.3 Logistic Regression Procedure

A viable alternative to IRT procedures that addresses some of its disadvantages is to use the logistic regression procedure. Like item response models, logistic regression procedures are also model based and can account for the continuous nature of ability (Swaminathan & Rogers, 1990). In addition, these procedures are able to accommodate small samples, are associated with well-accepted statistical tests of significance, and condition on observed, rather than latent, scores (Bennett, Rock & Kaplan, 1987; Hills, 1989; Swaminathan & Rogers, 1990). Also, logistic regression procedures are easier to use and understand than IRT-based procedures, and are readily available in standard statistical computer packages (Hills, 1989; Hosmer & Lemeshow, 1989). Compared to the Mantel-Haenszel procedure, the LR procedure has the advantages that non-uniform DIF can be detected, and conditioning can easily be extended to multiple variables.

The LR model is based on the equation

$$P(u=1 | X) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} \quad (2.5)$$

where u is the response to the item, X is the observed ability of the individual, β_0 is the intercept parameter, and β_1 is the slope parameter. The use of the LR model to detect DIF was proposed by Swaminathan and Rogers (1990) as it takes into account the continuous nature of the ability scale, is able to detect uniform and nonuniform DIF, and enables the incorporation of two or more covariates into the equation. The equation:

$$P(u_{ij} = 1 | X_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}X_{1j})}}{1 + e^{(\beta_{0j} + \beta_{1j}X_{1j})}} \quad (2.6)$$

where $i = 1, \dots, n_j$ (representing the examinees) and $j = 1, 2$ (representing the number of groups), was used by the authors to identify any DIF for the two groups of interest. If $\beta_{01} = \beta_{02}$ (i.e. intercepts are equal) and $\beta_{11} = \beta_{12}$ (i.e. slopes are equal) the logistic curves of the item for the two groups of interest are the same, and thus the item does not display any DIF. If, however $\beta_{01} = \beta_{02}$ and $\beta_{11} \neq \beta_{12}$, the curves are parallel but not coincident and hence uniform DIF may be inferred. On the other hand, if $\beta_{11} \neq \beta_{12}$, the curves are not parallel and thus non-uniform DIF exists, irrespective of whether the intercepts (β_{01}, β_{02}) are equal or not.

An alternative method of representing the model proposed by the authors is:

$$P(u=1) = \frac{e^z}{1+e^z} \quad (2.7)$$

where $z = \tau_0 + \tau_1 X + \tau_2 g + \tau_3 X_g$. Here $g = 1$ if the examinee is a member of group 1, and $g = 0$ if the examinee is a member of group 2, X_g is the product of g and θ , τ_2 is the group difference and τ_3 is the interaction

between group and ability. That is, $\tau_2 = \beta_{01} - \beta_{02}$ and $\tau_3 = \beta_{11} - \beta_{12}$. An item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$; and nonuniform DIF if $\tau_3 \neq 0$ irrespective of τ_2 . These hypothesis (i.e. $\tau_2 = 0$ and $\tau_3 = 0$) can be simultaneously tested as part of the null hypothesis $H_0: C_\tau = 0$ against $H_A: C_\tau \neq 0$, using the matrix:

$$C = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \quad (2.8)$$

when comparing three groups (i.e. 1 v 2 and 1 v 3). The statistic for testing the hypothesis is:

$$\chi^2 = \hat{\tau}'C'(C \sum C')^{-1}C\hat{\tau}' \quad (2.9)$$

which has a χ^2 with 2 degrees of freedom. The hypothesis of no DIF, that is $H_0: C\tau = 0$, is rejected when Equation 2.8 is greater than $\chi^2_{\alpha;2}$.

In the context of multiple group comparisons, Equation 2.6 enables the identification of DIF in the respective target groups while taking into account the influence or effect of other groups in the study. This can easily be done by specifying an appropriate matrix to compare each target group to the rest of the groups in the study. Also, an additional advantage of the logistic regression model is that it is possible to include other relevant factors into the model so as to identify possible explanations for DIF (Mazor, Kanjee & Clauser, in press). That is, if additional information about a group of examinees is available, this information could be included as covariates in the equation and used to help identify and/or explain possible reasons for DIF. This is especially relevant for cross-cultural comparisons where additional information or explanations about different groups could greatly improve our knowledge and understanding of these groups.

2.6 Questions That This Study Addresses

From the above discussion of the various judgmental and statistical methods of assessing equivalence and identifying DIF, it is evident that a great deal more research and information is still required before the proper application of these methods is mastered, and the questions that these comparisons raise are adequately addressed. This is especially true for cross-cultural/language studies. Some of the specific problems related to these applications include the development of methods and techniques for addressing: (1) cross-cultural comparisons involving multiple language groups, (2) small or modest group sample sizes for DIF and equivalence studies, (3) comparisons between groups with [vastly] different ability distributions, (4) issues of multidimensionality in data, (5) the use of instruments from which polytomous data are derived (that is item/question format types that include a mixture of objective and subjective items/questions), and (6) the development of international standards and guidelines for conducting cross-cultural/language comparisons, that is the translation, adaptation, administration and interpretation of testing instruments, as well as the reporting and utilization of scores (Hambleton, 1993).

In this study, only the detection of DIF in multiple group comparison was addressed. Specifically, the problem of small sample sizes was investigated. In this respect, both the logistic regression and pseudo-IRT procedures, discussed above, appear to provide a viable solution. The applications of these procedures can be extended to assess DIF in multiple groups by combining the data and/or modifying the model used so as to accommodate three or more groups. However, data

regarding the performance of these procedures under different conditions are not generally available. Thus, it was the purpose of this study to:

1. assess the viability of using the PIRT and LR procedures to simultaneously detect DIF in multiple groups, and
2. investigate the impact of small sample sizes and differences in underlying ability distributions on the DIF statistic when multiple groups are compared using the PIRT and LR procedures.

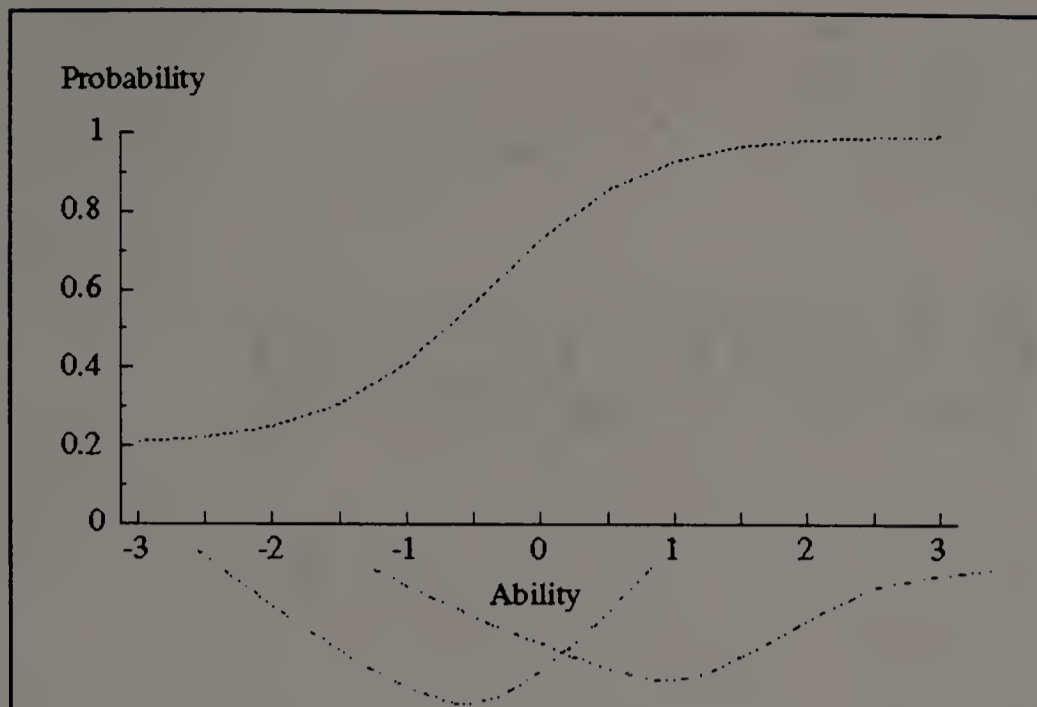


Figure 2.1

An ICC and ability distributions for two groups of examinees.

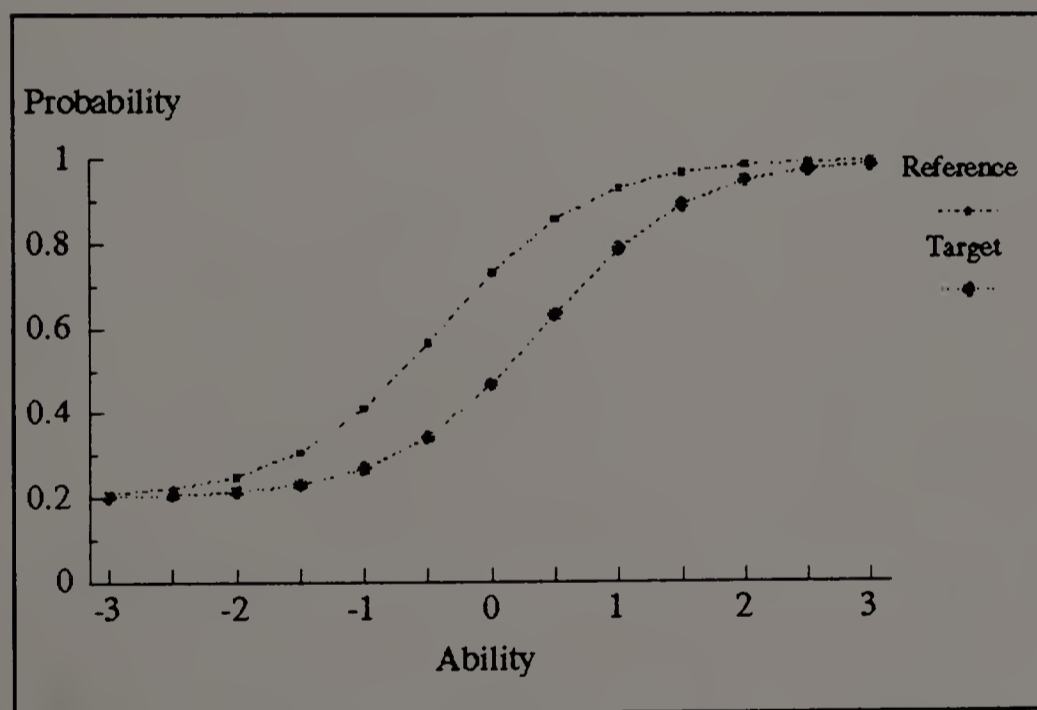


Figure 2.2

ICCs showing uniform DIF between target and reference groups

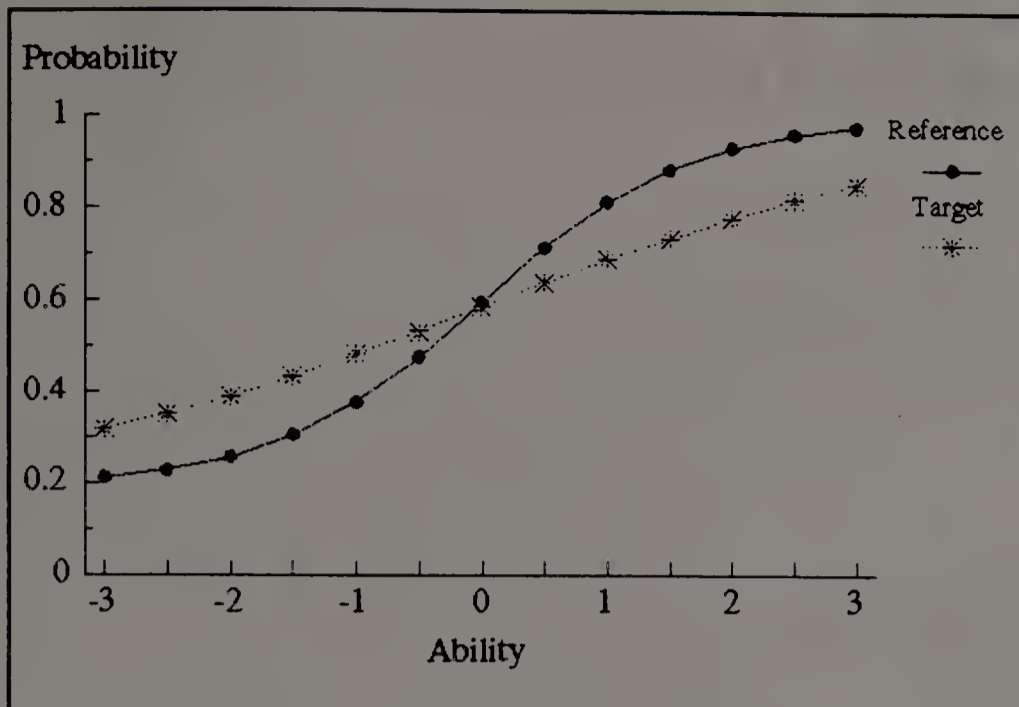


Figure 2.3

ICCs showing non-uniform DIF between target and reference groups

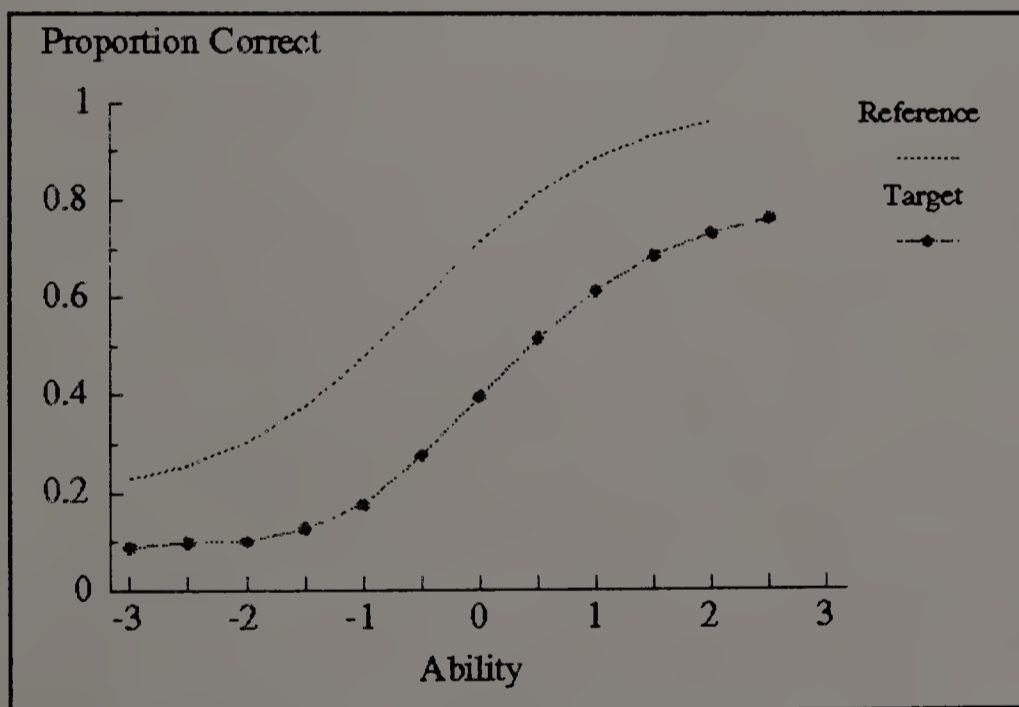


Figure 2.4

Observed and expected proportion correct as a function of ability for an item with large negative differences

CHAPTER 3

DETECTION OF DIF IN MULTIPLE GROUPS

3.1 Introduction

An important step in applying IRT to test data is the estimation of item and ability parameters that characterize the chosen item response model (Hambleton, Swaminathan & Rogers, 1991). Since only responses to test data are available and both the ability and item parameters are unknown, estimation procedures are used to determine the theta value for each examinee and the item parameters from the responses. A standard estimation procedure is to use the likelihood function for a set of item responses (Traub & Lam, 1985). However, to estimate item and ability parameters using these likelihood procedures, relatively large sample sizes as well as reasonably long tests are required, especially for the three parameter IRT model. Drasgow (1989) noted that Lord (1968) suggested samples of 1000 examinees and tests of 50 items for adequate estimation. Gifford (1983) also noted that these procedures require extremely large numbers of examinees and long tests for adequate estimation. This restriction severely constrains the use of these techniques in practice, for example, DIF studies.

A major problem in DIF studies is that of small sample sizes in the target group (Linn, 1993; Linn & Harnisch, 1981; Shepard, Camilli & Williams, 1985; Zieky, 1993). This problem is especially relevant to IRT-based procedures for detecting DIF. For example, sample sizes as large as 500 or more may be needed when the three parameter model is used (Hambleton & Swaminathan, 1985; Linn & Harnisch, 1981). When sample sizes are small, it is not unusual to exclude cultural

or language groups from any analysis. Excluding groups from any DIF analysis would affect the validity of the resulting test scores for the groups concerned if DIF items went undetected. Another disadvantage is that it is highly likely that valuable information is lost as well.

The approach proposed by Linn and Harnisch (1981) appears to provide a solution to overcoming the problem of small sample sizes in the target group for an IRT-based analysis. That is, item responses of the entire sample of examinees are used to obtain the item parameter estimates. This is especially relevant when three or more groups are compared. One approach is to combine the item responses of all the examinees as if they belong to one single large group, regardless of group membership, and then estimate the item and ability parameters. A second approach would be to first obtain the item parameter estimates from the largest group, then treat these parameter estimates as fixed to obtain the ability estimates for the target group(s). A third approach would be to first obtain the item parameter estimates from the total sample of examinees that excluded the target group, treat these item parameter estimates as fixed to obtain the ability parameter estimates for the target group(s).

In all three approaches, since only ability parameter estimates are obtained for the target groups, it is no longer possible to identify the existence of DIF by comparing ICCs. Instead, for each item the difference between the observed and predicted proportion correct (assuming the ICCs to be true for the target group) is calculated for every target group (Linn & Harnisch, 1981). When the pattern of residuals (observed - predicted performance) are not judged to be random via a chi-

square test or some other procedure, the item is identified as DIF. When sample sizes are increased by combining examinee item responses, the resulting IRT estimates obtained are more stable and random errors are reduced. However, the data samples of the different language/cultural groups could contain unknown levels of DIF. Thus combining data to increase sample sizes may result in an increase in systematic errors, and important DIF may go undetected, or at least this DIF is harder to detect.

3.2 Pseudo-IRT Estimation Procedures

To facilitate discussion regarding the different pseudo-IRT estimation procedures used, referred to as PIRT 1, PIRT 2 and PIRT 3, the following terms will be used: (1) target group (TRG) which refers to the specific group under investigation, (2) total group (TOT) which includes the sum of examinees across the different groups in the study (including the target group), and (3) adjusted group (ADJ) which refers to the combined groups excluding the (target) group under investigation, that is $ADJ = TOT - TRG$.

3.2.1 Pseudo-IRT Procedure 1 (PIRT 1)

In this procedure, item and ability parameter estimates are obtained for the total group only. That is, the item and ability parameters are only estimated once for the entire sample of examinees. Instead of comparing ICCs which is common in DIF studies, the Linn-Harnisch (1981) method was used to determine the existence of any DIF. An item is flagged as DIF if, for any target group, significant

differences between the observed and predicted proportion correct are found. That is, if the target group performance on the item is consistent with the best fitting ICC, the assumption is that there is no reason to suspect DIF.

The major advantage of this procedure is that item and ability parameter estimates are obtained only once and the item parameter estimates are based on the largest possible sample which is available. However, the disadvantage is that the ability estimates of the target groups are obtained from the total group based item statistics that *include* examinee item responses of the target group. Hence, the ability estimates of the two groups are no longer independent. If DIF does exist in the target group, the magnitude of the DIF indices would be deflated (Linn & Harnisch, 1981), decreasing the likelihood of detecting DIF (Ellis & Kimmel, 1992). Thus, it is possible that items exhibiting DIF would not be detected (i.e., type II errors would be increased). To overcome this problem, the PIRT procedure 2 could be used.

3.2.2 Pseudo-IRT Procedure 2 (PIRT 2)

In this procedure, item parameter estimates are first obtained from a predetermined (fixed) reference group, usually the largest group. The item parameter estimates are then treated as fixed to obtain the ability estimates for each target group. An item is studied for DIF by comparing the observed proportion correct at various ability levels to the expected performance using the reference group ICC. The advantage of this procedure is that the ICC is based solely on the reference group performance. Thus comparison between the target and reference

groups are not contaminated. The disadvantages are (1) since multiple estimations are required, this procedure could prove time consuming and more costly to use than estimation procedure 1, and (2) when the sample size of the reference group is not very large, the resulting DIF indices calculated may be inaccurate. In order to account for small sample sizes, PIRT procedure 3 can be used.

3.2.3 Pseudo-IRT Procedure 3 (PIRT 3)

In order to obtain more stable item parameter estimates, examinee item responses in this procedure are combined to increase the sample sizes such that for every group a corresponding reference group is created, called an adjusted group, which consists of examinees in all of the other groups. This procedure differs from procedure 1 as the item parameters estimates for every target group are obtained from a corresponding adjusted group that excludes examinee item responses from the target group of interest. These estimates are then treated as fixed to obtain the ability parameter estimates for the respective target groups. For example, the item parameter estimates are first obtained from the combined samples of groups 1 and 2 (that is adjusted group 12). These estimates are then treated as fixed in order to obtain the ability parameter estimates for group 3. An item is identified as DIF by comparing the expected and observed proportion correct in group 3 compared to the ICCs obtained for groups 1 and 2. The advantage of this procedure is that (1) the parameters estimated from the target and adjusted groups can be considered as independent, and (2) the resulting estimates are more stable as increased sample sizes are used in item parameter estimation. However, the disadvantage is that

multiple comparisons are required, which could prove costly and time consuming. Also, by combining data from the different groups, any DIF that exists in these groups may contaminate the data, and thus the detection of the DIF indices in the target group may be affected.

All three PIRT procedures are especially relevant for multiple groups comparisons, especially when one (or more) group(s) consists of small sample sizes. However, the major disadvantage is that for PIRT 1, the ability estimates of the target group are 'contaminated' since the item parameter estimates are based on responses that include members of the target group. For PIRT 2 and PIRT 3, multiple estimations, which are time consuming and costly to use are required. In addition, PIRT 2 cannot be used if none of the groups constitutes a large enough sample, while the problem with PIRT 3 is that the groups that are combined may themselves contain DIF, and thus the resulting estimates may be 'contaminated'. An alternative approach to the detection of DIF in multiple groups, that appears to be able to account for some of the disadvantages of the PIRT procedures, at least in theory, is to use the logistic regression procedure (LR). The next section notes some of the advantages of the LR procedure and presents two approaches of detecting DIF in multiple groups.

3.3 Logistic Regression Estimation Procedures

In the LR procedure, DIF in multiple groups is assessed by using a single logistic regression model to (1) estimate the item parameters, and (2) determine whether any significant differences exists between the target and adjusted groups in

their item performance. Unlike IRT procedures, it is possible to obtain stable item parameter estimates even when sample sizes are relatively small (Bennett et al., 1987; Swaminathan & Rogers, 1990), and estimation of the item parameters is relatively simple and straightforward. Typically, the estimates are obtained using the maximum likelihood method (Hosmer & Lemeshow, 1989). Logistic regression procedures are easier to understand and work with, and are readily available in standard statistical packages (Hills, 1989). DIF is assessed by comparing the logistic curves of test items for the groups under investigation. Items are flagged as DIF using the chi-square statistic to determine significant differences.

Two estimation procedures were used in the study. For both procedures, a single model was applied to detect DIF. In the first estimation procedure (denoted LR 1), each group was compared with each other, while in the second estimation procedure (denoted LR 2), each target group was compared to a corresponding adjusted group. The advantage of LR 1 (comparable to PIRT 2) is that the logistic curves which are compared were totally independent. However, the disadvantage is that multiple pairwise comparisons are required, which could become complicated to interpret (as lots of items being flagged but flagged for the various group combinations), and time consuming. For LR 2 (comparable to PIRT 3) the advantage is that data from different groups can be combined to increase sample sizes. However, the effect of this is that the estimates can become contaminated if data from other groups contain any DIF, and thus resulting DIF indices may be inaccurate.

3.4 Purpose and Research Questions

The previous sections outlined some potential advantages as well as problems when DIF is identified in multiple groups using IRT, and noted the use of LR procedures as a viable alternative. Both the logistic regression and pseudo-IRT procedures discussed above appear to provide viable alternatives to assessing DIF in multiple groups. The applications of these procedures can be extended to assess DIF in multiple groups by combining the data and/or modifying the model used so as to accommodate three or more groups. However, data regarding the performance of these procedures under different conditions are not generally available. It was thus the purpose of this study to determine:

1. the viability of extending applications of the pseudo-IRT and LR procedures to detect DIF in multiple groups,
2. which of the pseudo-IRT estimation procedures yield more accurate and reliable results when detecting DIF in multiple groups, and
3. the accuracy and reliability of the pseudo-IRT or LR procedures to detect DIF when multiple groups are compared.

This study was carried out using simulated procedures because data could be manipulated so that information of the amount and type of DIF as well as the location of the DIF items could be controlled for. In addition, the sample sizes and mean ability distributions of groups as well as the number of groups compared could also be controlled.

3.5 Method

3.5.1 Description of Data

The study was conducted on data simulated to fit a unidimensional three parameter model, using the computer program DATAGEN (Hambleton & Rovinelli, 1973). Data were simulated for three groups only as: (1) techniques that are applicable to three groups, can readily be adapted for use with more groups, and (2) to facilitate the application of statistical techniques and analysis of data. However, it is acknowledged that in practice, the number of groups in typical cross-cultural studies can range anywhere between three and ten or more (Ellis, 1991; Lapointe, Mead, & Phillips, 1989; Van der Berg, 1993, personal communication).

The data were simulated to represent examinee item responses for groups denoted GR1, GR2, and GR3. In addition, a total group, denoted TOT, and three adjusted groups, denoted ADJ12, ADJ23 and ADJ13, were created. TOT contained the combined examinee item responses across all the groups, irrespective of group membership, while ADJ12 excluded examinee item responses of GR3, ADJ23 excluded examinee item responses of GR1, and ADJ13 excluded examinee item responses of GR2.

Sample sizes were set at 1000 respondents per group so as to obtain stable item parameter estimates (Drasgow, 1989; Hambleton & Swaminathan, 1985). The ability distribution for all groups was normally distributed with mean 0 and standard deviation 1.0. A test length of sixty items was used as this is both within the range of typical standardized test lengths and long enough to reduce any instability that can occur with the results for shorter tests (Clauser, 1993). It also allowed for a

sufficient number of DIF items to be included in the test without the percent of DIF items in the test being excessive.

The percentage of items simulated with DIF for each group was 10%, an amount not unreasonable to find in practical testing situations (Mazor, Clauser, & Hambleton, 1991). DIF was simulated in both groups 2 and 3 using group 1 as a reference. Between groups 2 and 3 combined, the amount of DIF simulated was 20%. The amount of DIF (uniform vs non-uniform) was the same, however the specific items for which DIF was simulated differed. The first six items of group 2 and the last six items of group 3 exhibited DIF (See Table 3.1). The DIF items were also simulated to exhibit 50% uniform and non-uniform DIF. Uniform DIF exists when there is no interaction between ability level and group membership, while for non-uniform DIF there is an interaction between group membership and ability levels.

To simulate items to exhibit DIF, the differences were quantified in terms of the area between the curves for any two groups. The item parameter values were chosen to exhibit both uniform and non-uniform DIF. The a -values for uniform bias were set to 1.0 (reflecting moderate discrimination) while the DIF effect sizes (defined as the area between the ICCs) ranged from 0.4 to 0.8 (see Figures 3.1, 3.2, and 3.3), representing low to high DIF values. For non-uniform DIF, all a_1 -values were set at .6 and a_2 -values at 1.4 so as to obtain a high DIF size of .8. These values were selected so as to increase the probability of detecting the non-uniform DIF items. The b -values selected were set to -1.0, 0.0 and 1.0 (see Figures 3.4, 3.5, and 3.6) to represent low, medium and high difficulty items (Swaminathan &

Rogers, 1991). Table 3.1 shows the difficulty and discrimination values for the simulated DIF items, as well as the item numbers. Half of all the DIF items simulated for each group exhibited non-uniform DIF. In order to create conditions as close as possible to those found in actual practice, item parameters for all the non-DIF items were taken from parameter estimates of actual test items (item parameter values from one of the 1985 administrations of the Graduate Management Admissions Test).

3.5.2 Procedure

In this section, the specific DIF detecting procedures based on (1) PIRT, and (2) LR methods will be discussed.

3.5.2.1 Pseudo-IRT Procedure

Step 1 - Select the item response model. The three parameter logistic model was used because the data were assumed to come from the administration of a multiple choice item test.

Step 2 - Obtain the item and/or ability parameter estimates from the different examinee item response data sets. The program BILOG V3.6 (Mislevy & Bock, 1990) was used to estimate the parameters for the data sets. First, item and ability estimates were obtained for the total group, TOT. Second, GR1 was used as the fixed reference group to obtain the item parameter estimates, which were then treated as fixed to obtain the ability parameter estimates for GR2 and GR3. Third, for each target-adjusted group pair, item parameter estimates obtained from the adjusted group were treated as fixed in order to estimate the ability parameter

estimates for the respective target groups. For example, for the GR2-ADJ13 pair, item parameter estimates were first obtained from the examinee item responses in ADJ13. These item parameter estimates were then treated as fixed to obtain the ability estimates for GR2. This process was conducted three times, once for each of the different target-adjusted group pairs (GR1-ADJ23, GR2-ADJ13, and GR3-ADJ12).

Step 3 - Based on their ability scores, members of each target group were assigned into 10 score categories defined according to the: (1) equal N, (2) equal θ , and (3) equal probability approach (these approaches were discussed in Chapter 2). It must be noted that the θ values used to define the score regions included items that contained DIF.

Step 4 - Within each score category, Z_{iq} was computed (see equation 2.2). The chi-square statistic with 8 degrees of freedom was used to detect any significant differences. This figure was obtained by subtracting the number of item parameters estimated, that is 2, from the number of score categories used, that is 10 (Yen, 1981). Even though the three-parameter IRT model was used, only two parameters were estimated since the c-parameters were fixed at .20. All items that exhibited any significance differences between observed and expected item performance at the .01 significance level were flagged.

Step 5 - Twenty replications of the entire procedure, including the data simulation and parameter estimation, were conducted. All items, DIF and non-DIF (type 1 errors), that were flagged as DIF were recorded.

3.5.2.2 Logistic Regression Procedure

The logistic regression model used for comparing the target and adjusted groups is:

$$P(u_{ij} = 1 | X_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j} X_{1j})}}{1 + e^{(\beta_{0j} + \beta_{1j} X_{1j})}} \quad (3.1)$$

where u = the item response, X = is the observed ability of an individual, i represents the examinee, j represents the group, β_0 is the intercept parameter, and β_1 is the slope parameter. To detect DIF, separate equations were specified for the groups to be compared, and all items that exhibited significant differences at the .01 level, as measured by the chi-square statistic, were recorded. Since two different procedures were used, LR 1 and LR 2, the target groups were compared to: (1) group 1, selected to represent a fixed reference group for LR 1, and (2) the respective adjusted group, that is Equation 2 was used for every TRG-ADJ pair compared for LR 2. The former procedure (LR 1) is directly comparable to the PIRT 2, while the latter procedure (LR 2) is comparable to PIRT 3.

3.6 Results

Tables 3.2 and 3.3 summarize the results obtained for the pseudo-IRT procedures 1, 2 and 3 (denoted PIRT 1, PIRT 2 and PIRT 3) and logistic regression procedures 1 and 2 (denoted LR 1 and LR 2), respectively. The percentage of DIF items correctly flagged (detection rate) as well as the percentage of non-DIF items flagged (false positive errors) are reported. Because the three methods of dividing target group members into score categories yielded similar results, only the results

based on the Equal N procedure (total number of target group examinees are equally divided into 10 score categories) are reported for the PIRT procedures. All percentages were calculated over 20 replications, and all values reported were rounded off to their nearest whole. A significance level of .01 was used for testing the significance of the chi-square statistic for the individual items, thus an acceptable rate for the percentage of false positive errors detected would be approximately 1%. All groups had sample sizes of 1000, with the mean ability distributions and standard deviations set to 0 and 1, respectively.

Table 3.2 contains the results of the PIRT procedures. Results are reported for target groups GR2 and GR3 (each contained 10% DIF items), as well as for GR1 (that contained no DIF). In comparison to GR1, DIF items detected in GR2 and GR3 function against members in GR2 and GR3, respectively. For PIRT 2, the detection rate and false positive errors are reported for each target group, GR2 and GR3 that were compared to GR1. For procedures PIRT 1 and PIRT 3, detection rates were reported such that the contribution of DIF items from GR2 and GR3 are noted as well since in both these procedures the data from the different groups were combined. Therefore it was expected that the DIF items that were located in GR2 and GR3 would also be flagged, no matter which target group was under investigation. Also, by ignoring these items, the number of items incorrectly flagged as DIF would be erroneously inflated as actual DIF items would then be counted as false positive errors. When DIF is investigated in GR2, GR3 becomes the non-target group, while when GR3 is investigated, GR2 becomes the non-target group.

In addition to target GR2 and GR3, DIF analysis was also conducted in GR1, as this analysis would reveal items that function in favor of GR1. Schmitt (1988) notes that in practice, the main interest with regard to DIF studies has too often been on only those DIF items that function against any group. Since all DIF items that functioned against GR2 and GR3 were simulated using GR1 as a reference, DIF items detected in GR1 functioned in favor of members in GR1 relative to GR2 and GR3. Thus both GR2 and GR3 become non-target groups, and items favoring GR1 in comparison with GR2 and GR3 were expected to be detected.

When GR2 was under investigation, it was expected that DIF items 1 to 6 located in GRP 2 would be flagged. However, since the analysis was conducted on combined data, DIF items 55 to 60 (located in the non-target group GR3) would also be detected, even though GR3 was not under investigation. Similarly, for target GR3, DIF items located in the non-target group GR2, and for target GR1, items located in both GR2 and GR3 were expected to be flagged. For example, when GR2 was compared to TOT using PIRT 1 (see Table 3.2), 92% of the DIF items located in GR2 were correctly detected and 3% of the items were falsely flagged. In addition, since the estimates were based on combined data, 48% of the items showing DIF in GR3 against GR1 and GR2 were identified.

Similarly, results are reported for the logistic regression procedures in Table 3.3. For LR 1, the percentage of items correctly detected as well as the percentage of false positive errors are reported for each pairwise comparison: GR1 v GR2, GR1 v GR3, and GR2 v GR3. However, since the data were simulated such that both GR2 and GR3 contained DIF, results for the GR2 v GR3 comparison, are reported

such that the location of the DIF items are noted as well. For procedure LR 2, the results are reported similar to those for procedures PIRT 1 and PIRT 3.

All three PIRT procedures detected a significant amount of the DIF items in the respective target groups GR2 and GR3. For both PIRT 2 and PIRT 3, for which ability estimates for the target groups were obtained independent of the respective target groups, 100% of the DIF items in both target GR2 and GR3 were detected. In addition, approximately 70% of the DIF items for PIRT 3 were correctly identified in the non-target GR2 and GR3 respectively. For PIRT 1, for which the item and ability parameter estimates were obtained from the total group of examinees, 92% and 96% of the DIF items were detected for target GR2 and GR3 respectively. 48% of the DIF items in the non-target GR2 and GR3 were also correctly identified. As expected, only 1% of the items were flagged as DIF in GR1 using PIRT 2, since a significance level of .01 was used to test the significance of the chi-square statistic for the individual items.

The percentage of false positive errors obtained for all the PIRT procedures were relatively high. Both PIRT 2 and PIRT 3, which displayed high detection rates, had correspondingly high false positive errors as well. For PIRT 3, the false positive errors were 20 and 24%, for PIRT 2 it was approximately 11%, and for PIRT 1, it was approximately 4% in target GR2 and GR3, respectively. It appears that when the item and ability parameter estimates were obtained from the larger sample sizes (PIRT 1 with a total sample of 3000 examinees), the number of false positive errors was lower for each target group, even though the estimates in the target groups were contaminated by the responses from the other target groups.

Considering that only about 4% of the DIF items were not detected, which translates to less than 1 item per test, and that the false positive error was only 4%, which translate to approximately 2 items per test, PIRT 1 seems the procedure of choice in the conditions simulated. This is especially encouraging as PIRT 1 only requires a single analysis to obtain the item and ability parameter estimates. However, it must be noted that in this situation, the IRT estimates were obtained on large sample sizes and that the mean ability distributions of the groups compared were all equal.

For the LR 1(see Table 3.3), 95% and 98% of the DIF items were consistently flagged when target GR2 and GR3 (respectively) were compared to GR1. However, for LR 2, 100% of the DIF items were correctly identified in target GR2 and GR3 respectively. Also, since LR 2 involved combining data of the different groups, approximately 78% of the DIF items in the non-target GR3 (for the groups 13 v 2 comparison) and GR2 (for the groups 12 v 3 comparison) was also identified.

When GR1, was treated as the target group, items detected as DIF function in favor of GR1. About 24 and 35% of DIF items for PIRT 1, and 63% of the DIF items for PIRT 3 were flagged in target GR2 and GR3, respectively. However the percentage of false positive errors for PIRT 1 was only 4% as compared to 20% for PIRT 3. Similarly, for LR 2, when the combined responses of GR2 and GR3 were compared to GR1, 62% and 58% of DIF items were flagged in GR2 and GR3, respectively. The low detection rate for PIRT 1 was primarily because in the GR1 vs TOT comparison, only a third of the DIF items function differentially with

respect to GR1, while in the GR2 vs TOT or GR3 vs TOT comparisons, half of the DIF items function differentially with respect to GR2 and GR3, respectively. When the target group responses were excluded from the total group, PIRT 3 and LR 2, approximately two-thirds of DIF items were detected for PIRT 3 and LR 2. However, 33% of the DIF items were still undetected. The significance of this result is that detection rates are more accurate and higher when the existence of DIF is directly determined in the specific target group.

A very encouraging result with respect to LR 1 is that when GR2 and GR3 were directly compared to each other, 100% of the DIF items were flagged in both groups, and the false positive errors were only 1%. Comparing the DIF groups directly improved the detection rates by 5% for the groups 1 v 2 comparison and by 2% for the groups 1 v 3 comparison. However, in practice, prior knowledge of the existence of DIF in any group is not available and thus this comparison may not have any practical significance. That is, if multiple pairwise comparisons is the method of choice to detect DIF in three or more groups, and a single group is used as a reference, comparing every group to each other only improves the detection rate by approximately 3%. For a 60 item test, this translate to only one item. This increase may not be worth the time and effort required, especially if many groups are being studied.

Form the results it appears that the use of the LR procedures are preferable to the PIRT procedures, primarily because the false positive error rates were significantly lower. When a single reference group was used to detect DIF, (PIRT 2 and LR 1), even though the detection rate was slightly higher for the PIRT 2, the

false positive error rates were 10% and 11% in both target GR2 and GR3 respectively, as compared to only 1% and 2% in LR 1. When the target groups were compared to their respective adjusted groups, the detection rate were comparable in PIRT 3 and LR 2. However, the false positive errors for PIRT 3 was 20% as compared to 1% for LR 2.

The power of the simultaneous comparison procedures to detect DIF (PIRT 1, PIRT 3 and LR 2) was as high, and sometimes higher than that of the two-group comparisons (PIRT 2 and LR 1) in target GR2 and GR3. That is the simultaneous procedures, PIRT 1, PIRT 3 and LR 2, were just as reliable in detecting DIF as were the more commonly used (and accepted) two-group procedures, PIRT 2 and LR 1. Of the three simultaneous DIF detection procedures, the use of LR 2 is preferred as the false positive errors were the lowest, that is 1% compared to approximately 10% for PIRT 1 and 20% for PIRT 3.

In Table 3.4 and Table 3.5 the results for the different item types detected in target GR2 and GR3 are summarized for the PIRT and LR procedures, respectively. The item number, its location, a- and b-values, the DIF sizes as well as the detection rate for each DIF item is reported for each of the target groups GR1, GR2 and GR3. In addition, in those procedures where data from the different groups were combined (PIRT 1, PIRT 3 and LR 2), items that were flagged in the non-target GR3 (when GR2 was under investigation) and in non-target GR2 (when GR3 was under investigation) are presented as well. Last, items that functioned in favor of GR1 were also noted.

Since large sample sizes were used to obtain the parameter estimates, the detection rates for both uniform and non-uniform DIF items were expected to be high for both the PIRT and LR procedures. The detection rate for uniform DIF items were expected to increase as the DIF size increased, since items with greater DIF sizes would be easier to detect. Based on results reported by Rogers (1989), the detection rate for non-uniform DIF items (the DIF size was equally high for all the items) with moderate b-values (items 4 and 58) were expected to be higher than those with low (items 5 and 59) and high (items 6 and 60) b-values.

In Table 3.4 the detection rates for the specific item types are reported for the PIRT procedures. For uniform DIF items, the detection rate increased as a function of DIF size, as expected. For the non-uniform DIF items, the detection rates for the low difficulty (or easy) items were consistently higher than the moderate and high difficulty items, while the moderate difficulty items showed higher detection rates than the high difficulty items. This pattern is especially evident for PIRT 1 and for the non-target groups in PIRT 3. These results for the uniform DIF items are consistent with those reported by Rogers (1989). In addition, Clauser (1993) also found that the items with higher b-values were associated with lower detection rates. However, Clauser (1993) used the Mantel-Haenszel procedure in his study.

A similar pattern was detected for LR procedures. This is especially evident in non-target GR2 and GR3 for LR 2. For LR 1, in both the group 1 v 2 and group 1 v 3 comparisons, the uniform DIF items with the smallest DIF size and high difficulty non-uniform DIF items were harder to detect. For LR 2, 100% of

all the DIF items were detected in each target group GR2 and GR3. This is probably due to the increase in sample sizes as the parameter estimates were based on the combined data samples. However, the detection rate pattern noted for the PIRT procedures is much more consistent and clear in the non-target groups for LR 2.

Compared to LR 1, PIRT 2 had a higher detection rate for uniform DIF items with small DIF size and non-uniform DIF items with high b-value. The detection rate for both PIRT 3 and LR 2 was 100% for all item types in target GR2 and GR3, thus no comparison could be made between the different item types. However, in the non-target groups, the PIRT procedure had a higher detection rate for uniform DIF items, while the LR procedure had a higher detection rate for non-uniform DIF items. This finding indicates that when data are combined, LR procedures are better at detecting non-uniform DIF, while PIRT procedures are better at detecting uniform DIF.

3.7 Discussion

This study demonstrates that both pseudo-IRT and LR procedures are viable techniques for simultaneously detecting DIF in multiple groups. When sample sizes are large, and ability distributions of the groups are similar, LR 2 is the recommended procedure to simultaneously detect DIF in multiple groups. However, PIRT 1 also has adequate power to detect a significant number of the DIF items. The only disadvantage is that the number of false positive errors detected is slightly

higher than that of LR 2. The crucial point is that both IRT and LR techniques can be effectively used to simultaneously detect DIF in multiple groups.

This result has significance for multiple group comparisons. First, the standard definition of DIF that is relevant to two-group comparisons does not apply. That is, since DIF is a relative phenomenon, the existence of DIF in any one group can only be defined in reference to another group. For pairwise comparisons, this situation works well as each group serves as a reference for the other. However, when multiple groups are compared, the definition of DIF becomes unclear since pairwise and multiple group comparisons may identify different sets of items as DIF. Therefore, it is not certain which set of items exhibits DIF. This problem was noted by Ellis (1991), in a study comparing Americans, Germans and French students. Ellis (1991) found that items exhibiting DIF in the pairwise comparisons, may or may not exhibit DIF when compared to the total (combined) sample of examinees. Thus, the problem in multiple group studies becomes one of defining and identifying ("true") DIF items, the resolution of which has consequences for the interpretation of scores.

When the performance of three or more groups is compared, the definition of DIF must be revised so as to take into account the examinee responses of all groups in the study. Thus only those items, that are flagged as DIF when all three groups are simultaneously considered, need to be eliminated or revised. Given this, the definition of DIF for multiple group comparisons should read: "two or more versions of an item should be considered equivalent if members of same ability in each target and adjusted group have the same probability of success on the item."

Second, the performance of the different groups of interest can be directly compared since 'real DIF' from the data was eliminated and scores are equivalent. In a two-group comparison, the performance of different groups could only be compared indirectly, that is, in terms of a single reference group.

Third, the perception (and practice) of an existing hierarchy where the reference group is seen as "the standard" is eliminated. All groups are compared to what Ellis and Kimmel (1992) refer to as the "omnicultural composite" which includes all the groups in the study, and which is "truly" representative of any "standard".

Fourth, simultaneous DIF detection procedures work well when instruments are translated from more than one (base) language. For example, in their study, Ellis and Kimmel (1992) first translated parallel German and English versions of an instrument. From the German version, a French version developed, which was then back-translated into English. Final versions of the English, French and German versions were determined by comparing the two English versions (that is, the original and back-translated French version) as well as the German and French versions.

In the Ellis and Kimmel (1992) study, no single group was selected as the base group from which all instruments were to be translated, and which would then be used as a "standard" against which the existence of DIF would be determined. This is contrary to the procedures adopted in the Second International Mathematics and Science Study, where the performance of the U. S. participants was selected as the reference group in a number of DIF studies (Lapointe, Mead, & Phillips, 1989). That is, DIF studies were carried out by comparing performance in each

participating national group to the performance of the U. S. sample. Also, English was used as the 'base' language from which tests were translated into other languages. It could well be argued that the results could have been different had the basis of comparison been different, or if all the instruments were not translated from English only.

Fifth, the use of simultaneous DIF detection techniques noted eliminates the need for multiple pairwise comparisons, thus: (1) the type I errors are reduced, (2) the problem of "unmanageable" numbers of comparisons when many groups are compared is overcome (that is, for n groups a minimum of $n(n-1)/2$ pairwise comparisons are required), and (3) the percentage of items flagged as "true" DIF is reduced, since in pairwise comparisons it is possible that many different items will be flagged as DIF for each of the pairwise comparisons.

This study also has several limitations. First, these procedures are sample dependent as the definition of DIF depends on the specific sample of language/cultural etc. groups. Thus information derived from these studies cannot be used to compare performance of whatever construct measured across different studies. For example, it is entirely possible that items detected as DIF when French, Arabic and German speaking students are compared may not be the same as items which are flagged as DIF for comparisons between Swahili, Portuguese and Hindi speaking students, even if instruments were developed to include all these languages.

Second, this study was developed using relatively large samples. The item parameter estimates obtained from these large samples were stable, and thus more

reliable. In practice, such sample sizes are not readily available. The application of these studies must be tested against smaller sample groups to determine how they function under real life conditions. In the next chapter, the effect of small sample sizes on the PIRT and LR techniques are studied.

Table 3.1

Item Numbers and a- and b-values for Uniform and Non-Uniform DIF Items

Uniform DIF					Non-Uniform DIF				
Item Numbers		a- and b-values			Item Numbers		a- and b-values		
GRP2	GRP3	a	b1	b2	GRP2	GRP3	b	a1	a2
1	55	1.0	-0.2	0.2	4	58	0	0.6	1.4
2	56	1.0	-0.3	0.3	5	59	-1	0.6	1.4
3	57	1.0	-0.4	0.4	6	60	1	0.6	1.4

Table 3.2

Percentage of DIF Items Detected (over 20 replications) for Pseudo-IRT Procedures 1, 2 and 3
 (Sample size: GRP 1 = GRP 2 = GRP 3 = 1000)

Target group	PIRT 1 (TOT)		PIRT 2 (FR)		PIRT 3 (ADJ)	
	Detection rate DIF items in Grp 2	False positives Grp 3	Detection rate	False positives	Detection rate DIF items in Grp 2	False positives Grp 3
GRP 1	24	4	N/A	1	63	20
GRP 2	92	3	100	10	100	20
GRP 3	48	4	100	11	69	24

Table 3.3

Percentage of DIF Items Detected (over 20 replications) for LR procedures 1 and 2
 (Sample size: GRP 1 = GRP 2 = GRP 3 = 1000)

Comparisons	LR 1		LR 2	
	Detection Rate DIF items in Grp 2 Grp 3	False positives	Comparison	Detection rate DIF items in Grp 2 Grp 3
Groups 1 v 2	95	1	Groups 13 v 2	78
Groups 1 v 3	--	2	Groups 12 v 3	100
Groups 2 v 3	100	1	Groups 23 v 1	58
				2

Table 3.4
 Percentage of Item Types Detected Using Pseudo-IRT Procedures

No.	Origin Group	DIF ITEM			PIRT 1			PIRT 2			PIRT 3				
		a	b1	b2	Size	GR 1	GR 2	GR 3	Detection Rate	GR 1	GR 2	GR 3	Detection Rate	GR 1	GR 2
1	2	1	-.2	.2	.4	10	80	25	-	95	-	65	100	100	90
2	2	1	-.3	.3	.6	20	100	55	-	100	-	95	100	100	85
3	2	1	-.4	.4	.8	65	100	95	-	100	-	100	100	100	100
		<u>b</u>	<u>a1</u>	<u>a2</u>											
4	2	0	.6	1.4	.8	25	95	55	-	100	-	40	100	100	50
5	2	-1	.6	1.4	.8	25	100	40	-	100	-	55	100	100	75
6	2	1	.6	1.4	.8	0	75	15	-	100	-	20	100	100	15
		<u>a</u>	<u>b1</u>	<u>b2</u>											
55	3	1	-.2	.2	.4	5	30	85	-	-	100	75	90	100	100
56	3	1	-.3	.3	.6	45	75	100	-	-	100	90	100	100	100
57	3	1	-.4	.4	.8	70	85	100	-	-	100	100	100	100	100
		<u>b</u>	<u>a1</u>	<u>a2</u>											
58	3	0	.6	1.4	.8	20	30	95	-	-	100	40	45	100	100
59	3	-1	.6	1.4	.8	55	70	100	-	-	100	50	70	100	100
60	3	1	.6	1.4	.8	15	20	95	-	-	100	25	15	100	100

Table 3.5

Percentage of Item Types Detected Using Logistic Regression Procedures

No.	Origin	DIF ITEM		Size	LR 1			LR 2			
		a and b values	a		b	Group comparisons	1 v 2	1 v 3	2 v 3	13 v 2	12 v 3
1	2	1	.2	.4	85	-	100	100	100	65	25
2	2	1	.3	.6	100	-	100	100	100	95	85
3	2	1	.4	.8	100	-	100	100	100	100	100
		<u>b</u>	<u>a1</u>	<u>a2</u>							
4	2	0	1.4	.8	100	-	100	100	100	75	50
5	2	-1	1.4	.8	100	-	100	100	100	80	95
6	2	1	1.4	.8	85	-	100	100	100	45	20
		<u>a</u>	<u>b1</u>	<u>b2</u>							
55	3	1	.2	.4	-	95	100	70	100	100	20
56	3	1	.3	.6	-	100	100	100	100	100	80
57	3	1	.4	.8	-	100	100	100	100	100	95
		<u>b</u>	<u>a1</u>	<u>a2</u>							
58	3	0	1.4	.8	-	100	100	55	100	100	60
59	3	-1	1.4	.8	-	100	100	90	100	100	75
60	3	1	1.4	.8	-	90	100	55	100	100	20

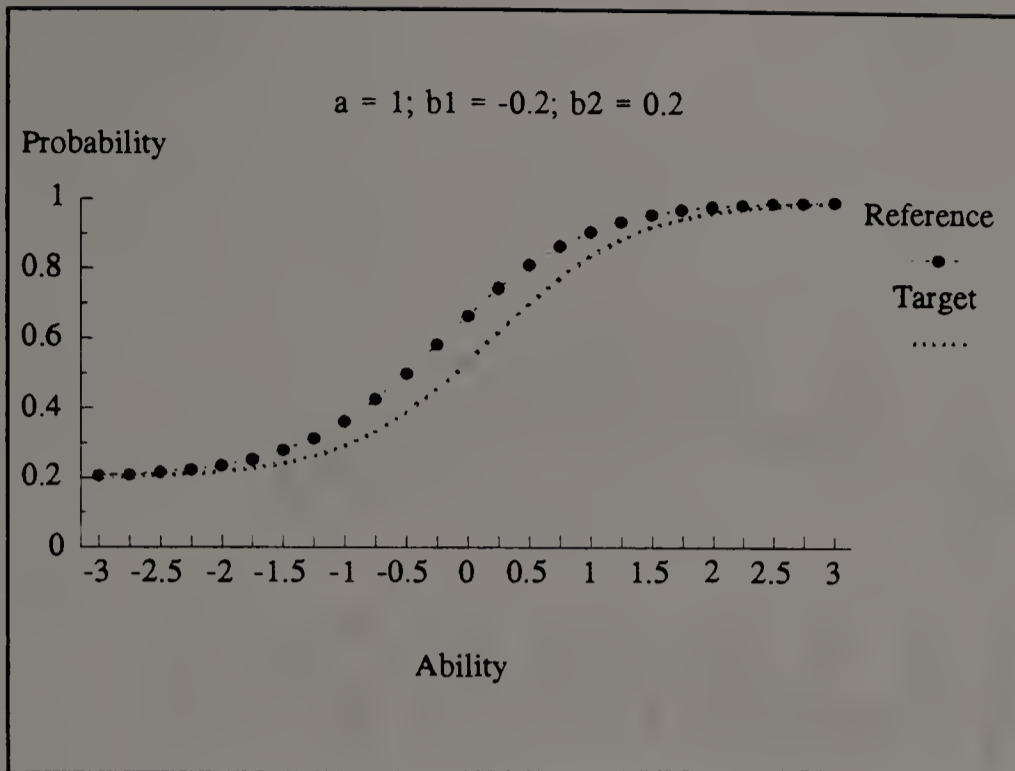


Figure 3.1

Uniform DIF item with area = 0.4

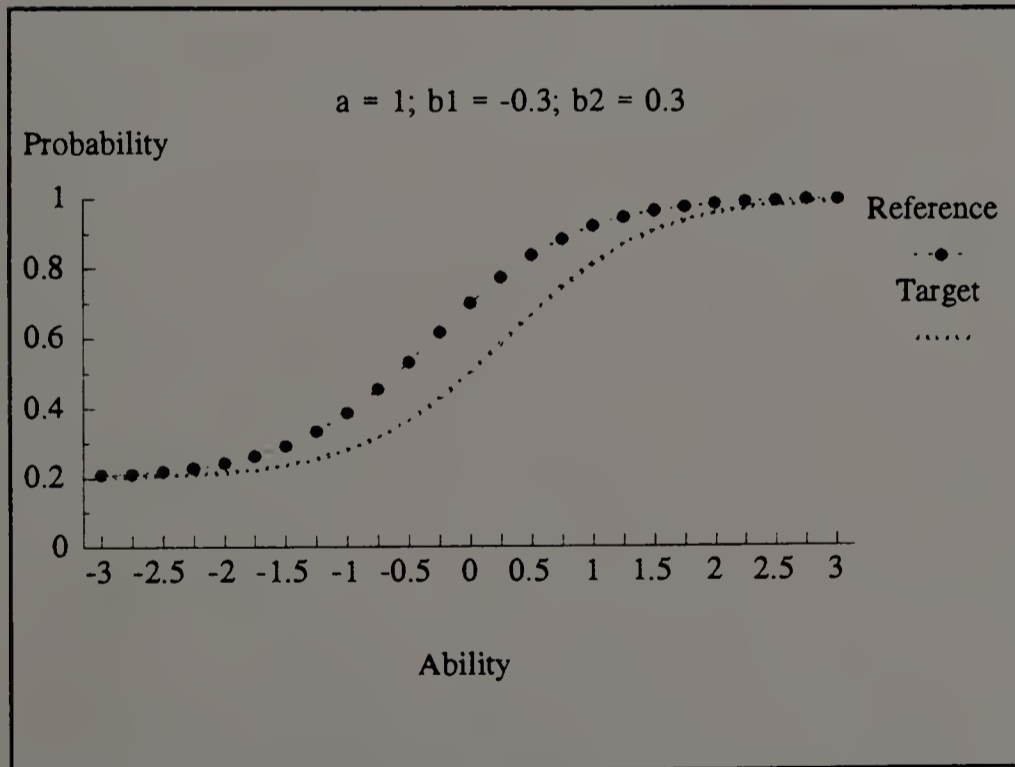


Figure 3.2

Uniform DIF item with area of 0.6

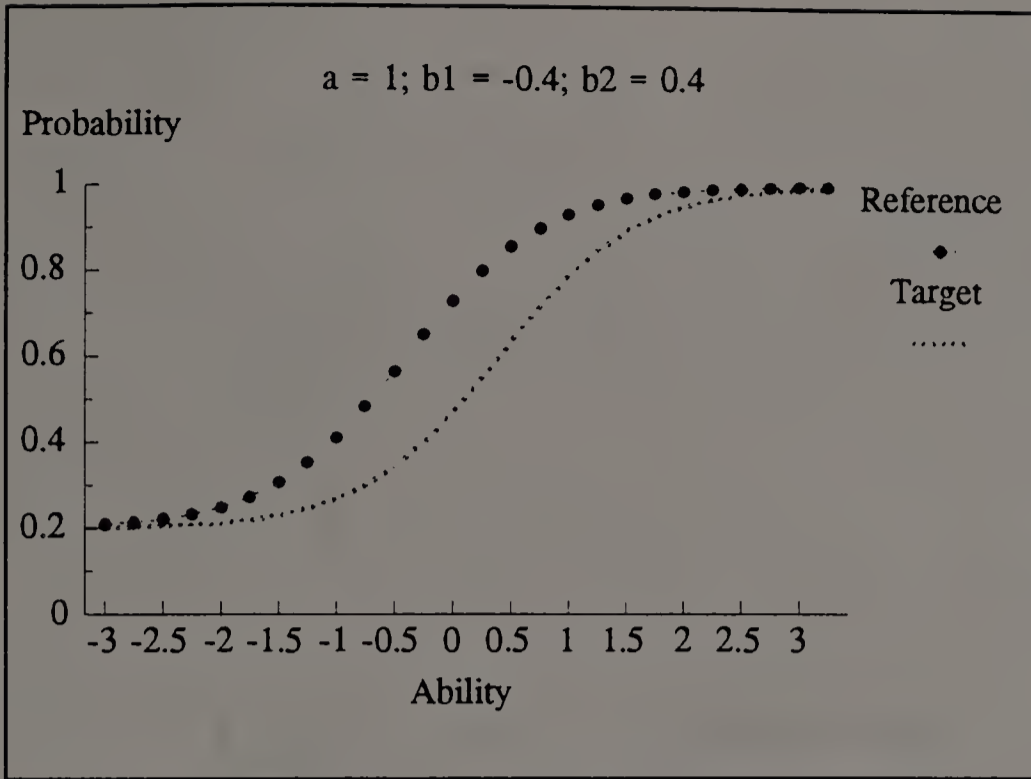


Figure 3.3

Uniform DIF item with area = 0.8

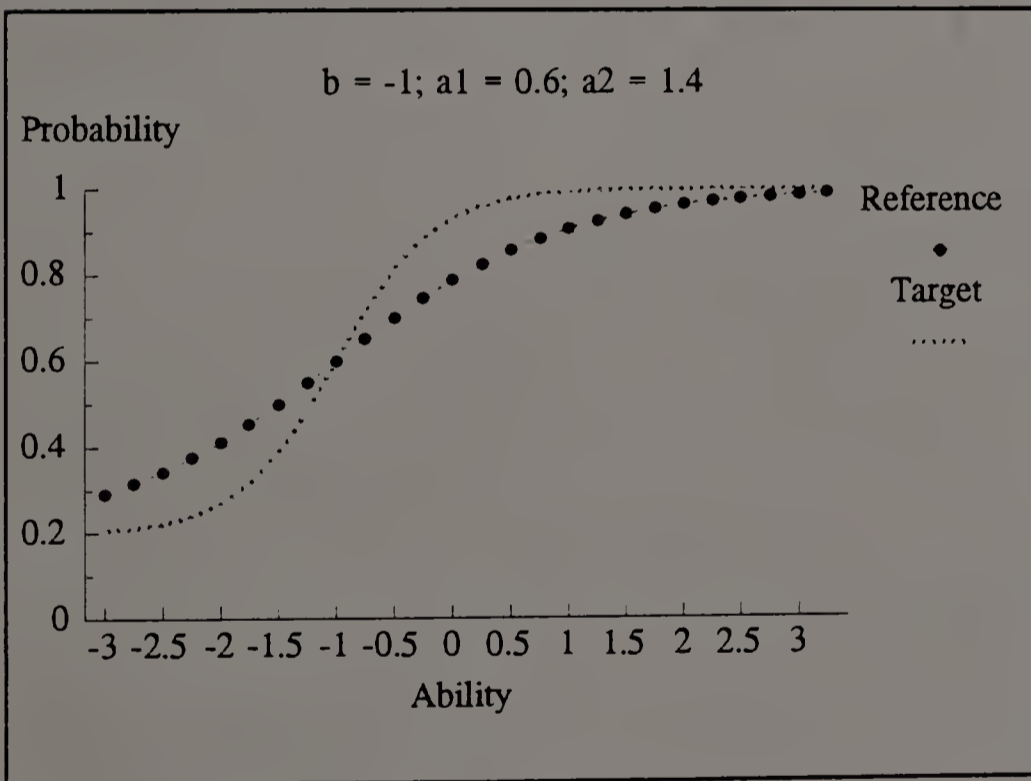


Figure 3.4

Low difficulty non-uniform DIF item (area = 0.8)

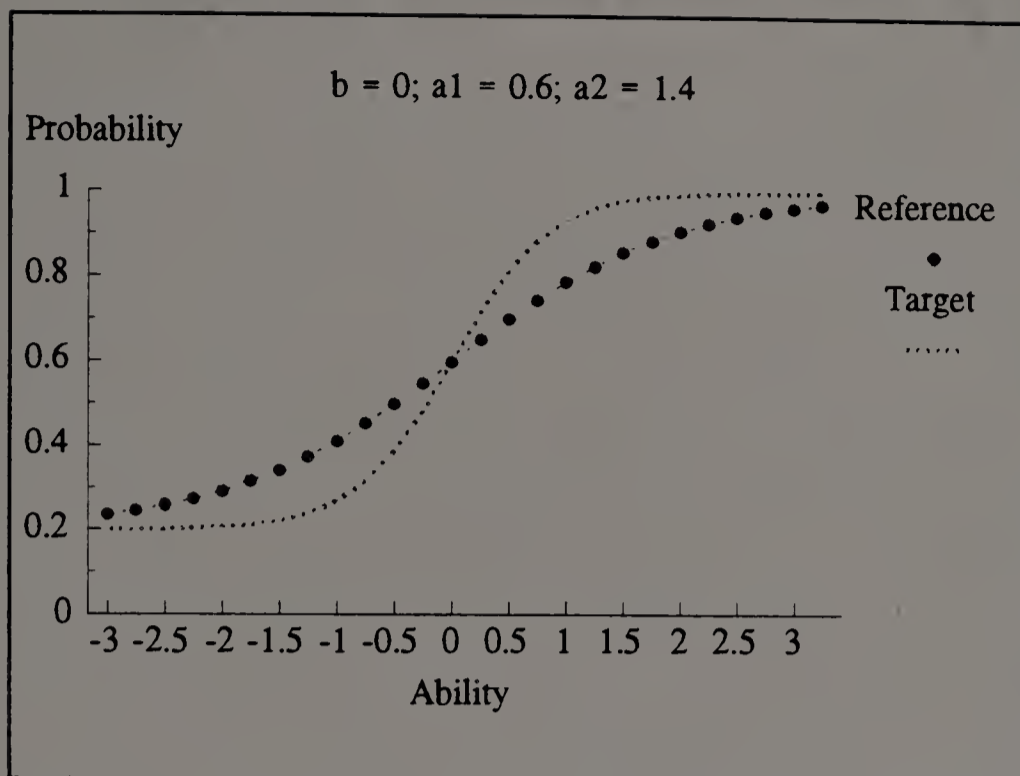


Figure 3.5

Moderate difficulty non-uniform DIF item (area = 0.8)

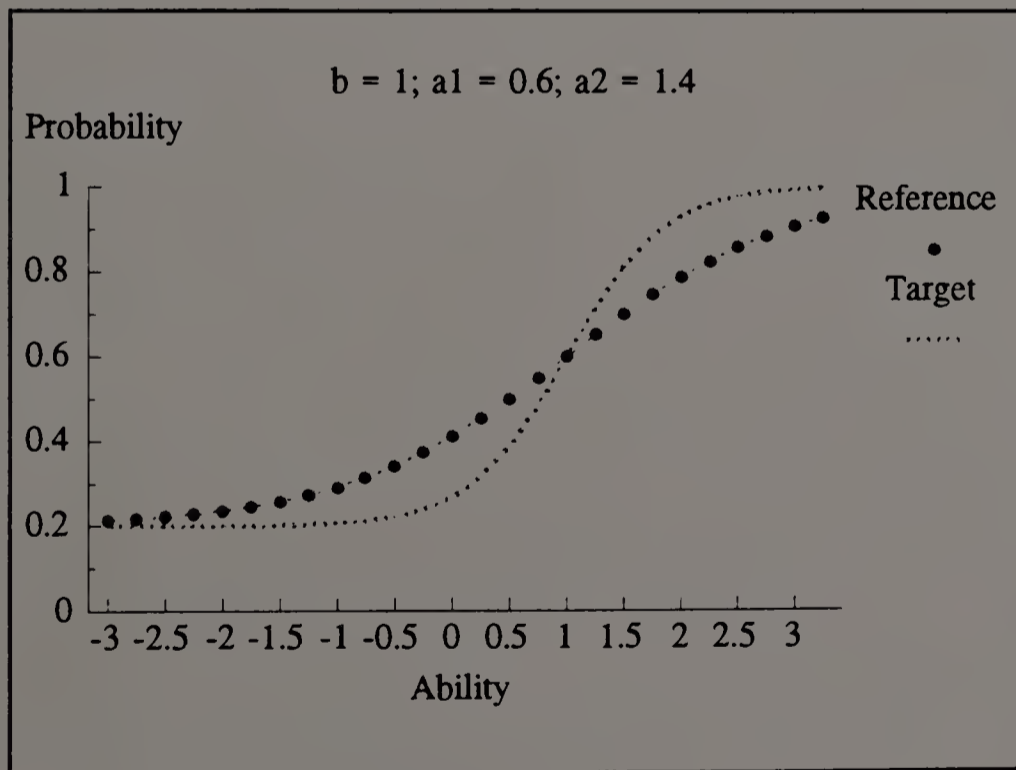


Figure 3.6

High difficulty non-uniform DIF item (area = 0.8)

CHAPTER 4

THE EFFECT OF SAMPLE SIZE AND ABILITY DISTRIBUTION ON THE DETECTION OF DIF IN MULTIPLE GROUPS

4.1 Purpose of the Investigation

The ability to detect DIF in instruments that are translated/adapted into different languages is a crucial aspect of any (large scale) cross-cultural/national study. A particular problem in this regard is the availability of adequate sample sizes in groups that represent minority (language or cultural) populations. The sample size (of target groups) is especially relevant to IRT procedures as relatively large sample sizes are required to obtain stable estimates (Linn & Harnisch, 1981). In practice it is not unusual to find target groups of less than 300 examinees, especially in small testing programs, thus limiting the use of IRT procedures (Hills, 1989; Parshall & Kromrey, 1992).

Several studies, some in the context of cross-language comparisons, that studied the impact of sample size on IRT DIF detection techniques produced some interesting findings. Budgell (1992) compared IRT and Mantel-Haenszel procedures for detecting DIF in English and French versions of numerical and reasoning tests using Raju's IRT signed and unsigned method. Sample sizes of a 1000 examinees were used for both groups of English and French examinees that took the 15-item numerical and 18-item reasoning tests. The means and standard deviations of both language groups were similar for each test. The results of Budgell's study supported the use of both Raju's sign and unsigned method, as well as the Mantel-Haenszel procedure for detecting DIF in translated instruments. However, when smaller

sample sizes were used (i.e. 199), Budgell found that only 12% of the items with significant DIF that were previously detected with samples of 1000, were identified.

Ellis (1991) conducted three studies using IRT to examine the measurement equivalence of three types of translated tests: ability tests, attitude surveys and personality tests. The tests were administered to American and Germans in English and German, respectively. Ellis (1991) compared the item parameter estimates and used Lord's chi-square as an index of DIF. She found that while the IRT analysis could not identify the source of DIF, IRT was certainly useful in identifying DIF items. However, Ellis (1991) noted that the samples sizes in all three studies were relatively small (ranging from approximately 200 to 300), and that using larger samples would result in more stable item parameter estimates, and thus produce more valid results.

Compared to IRT procedures, the use of LR procedures to identify DIF has been relatively infrequent. Swaminathan and Rogers (1991) compared the power and accuracy of the Mantel-Haenszel and LR procedures to detect uniform and non-uniform DIF. The authors found that for both procedures, DIF detection rates increased with increases in sample size. However, the effect of different ability distributions on the DIF statistic was not studied.

In another study, Mazor, Kanjee and Clauser (in press) used the LR procedure with real data samples of about 1000 to evaluate the effect of conditioning on two ability estimates on the detection of DIF in Chemistry and History tests. The authors found that the LR procedure could differentiate between DIF and multidimensional item impact when two ability estimates were incorporated into the

LR procedure. However, in this study as well, the impact of differing abilities was not studied as all the samples had similar underlying distributions.

Bennett, Rock and Kaplan (1987) used large samples in their study to identify DIF for handicapped examinees taking special extended-time administrations of the SAT. The authors found that the LR procedure detected notable instances of differential performance at the item level.

Typically in practice, different groups can be expected to have different ability distributions, and disadvantaged groups are more likely to have lower mean abilities than the majority or reference groups. Raju, Bode and Larson (1989) note that compared to the reference group (white North American students) examinees in the focal group (Black and Hispanic students) scored roughly one standard deviation below the reference group on a vocabulary test measuring basic skills. This is especially relevant in South Africa where the nature of the society is such that the non-dominant and different language groups also constitute the 'disadvantaged' groups¹. In DIF studies, where the focus is often on 'minority groups', it is important that the ability distributions of groups compared be taken into consideration.

The ability distribution of groups compared is especially relevant for those DIF detection techniques that compare ICCs and that condition on ability, for example, the Linn & Harnisch (1981) PIRT procedure. Often groups show sizeable

1

At least one of the two official languages are also the first language of White South Africans, who also enjoy the benefits of a vastly superior (by whatever indicator used) educational system as compared to Black South Africans.

differences in average ability, which is reflected by little overlap in ability distributions (Camilli & Shepard, 1994). Wainer (1993b) notes that when the differences in ability distributions are large, what appears to be a large area between the ICCs may only affect a small number of persons in the focal group (See Figure 4.1). For IRT methods this situation is problematic as the effective sample size is defined by this overlap because only conditional differences between groups are analyzed. Thus, an apparent large sample size may be insufficient for an analysis of DIF, while on the other hand, smaller sample sizes may be adequate for groups of nearly equivalent ability (Camilli & Shepard, 1994).

Mazor, Clauser and Hambleton (1991) studied the effect of sample size on the functioning of the Mantel-Haenszel (MH) statistic, and also compared samples with equal and unequal ability distributions. The results of their study indicated that (1) the percentage of DIF items correctly identified decreased markedly as the number of examinees decreased, and (2) the detection rates for equal and unequal ability distributions were very similar. However, Clauser (1993) found that the Mantel-Haenszel statistic identified fewer DIF items when the ability distributions between the two groups were unequal. Clauser further noted that when "groups of differing abilities are to be compared, it is probably advisable to be even more conservative and use larger samples" (p. 85).

In the studies presented, what is evident is that DIF detection techniques worked well with large samples. When sample sizes decreased, the number of DIF items detected decreased as well. Also, when ability differences between the groups were present, the number of items detected as DIF decreased. What is unclear

however, is the effect of the interaction of a lower mean ability distribution in the target group with a small sample size on the detection of DIF. In addition, all analyses in the studies noted were conducted using two groups only. Little data, if any, exists on detecting DIF when multiple groups are compared where DIF is assessed in all groups simultaneously. The effect of small sample sizes as well as groups differing in their average performance on detecting DIF simultaneously in multiple groups is not known. In this study, the effect of (1) small sample sizes, and (2) differing underlying abilities on DIF indices when multiple groups were compared using LR and IRT procedures was investigated.

4.2 Research Design

The data for this study were simulated to fit a unidimensional three parameter model, using the computer program DATAGEN (Hambleton & Rovinelli, 1973). The data were simulated to represent examinee item responses for three groups: GR1, GR2, and GR3. In addition, a total group, TOT, and three adjusted groups, ADJ12, ADJ23 and ADJ13, were also created. TOT contained the combined examinee item responses across all the groups, irrespective of group membership, while ADJ12 excluded examinee item responses of GR3, ADJ23 excluded examinee item responses of GR1, and ADJ13 excluded examinee item responses of GR2.

A test length of sixty items was used as this was both within the range of typical standardized tests and was long enough to reduce any instability that can occur with the results for shorter tests (Clauser, 1993). Also, this allowed for a sufficient number of DIF items to be included in the test. The percentage of items

simulated with DIF for each group was 10%. DIF was simulated in groups 2 and 3 using group 1 as a reference. Thus between groups 2 and 3, the amount of DIF was 20%. The first six items of group 2 (GR2) and the last six items of group 3 contained DIF.

The DIF items were also simulated to exhibit 50% uniform and non-uniform DIF. The a-values for uniform DIF were set to 1.0 while the DIF effect sizes (defined as the area between the ICC's) ranged from 0.4 to 0.8. For non-uniform DIF, all a-values differed by 0.8 while the b-values selected were 0.0, -1.0 and 1.0 (see Table 2.1). Half of the DIF items simulated in groups 2 and 3 exhibited non-uniform DIF. In order to create conditions as close as possible to those found in actual practice, item parameters for all the non-DIF items were taken from parameter estimates of actual test items (i.e., a 1985 administration of the Graduate Management Admissions Test).

4.2.1 Sample Size

In the first analysis, three groups were simulated with samples of 500, 300 and 100, respectively, for a total sample size of 900. A minimum sample of 100 was selected because this size may be found in practice and samples below 100 respondents do not provide stable item parameter estimates (Mazor, Clauser & Hambleton, 1991). The underlying abilities for all three groups were chosen to be normally distributed with mean 0 and standard deviation 1 (see Table 4.1).

4.2.2 Ability Distributions

In the second analysis, the ability distributions were varied as in practice it is more likely for the different cultural/ethnic groups to have lower mean abilities than the majority (or reference) group (Donoghue, Holland & Thayer, 1993; Raju, Bode & Larson, 1989). The ability distributions were normally distributed with a standard deviation of 1.0, while the mean abilities for groups 1 and 2 were set to 0 and the mean ability for group 3 was set to -1.0 (see Table 4.1).

4.2.3 Description of DIF Procedures

The LR and IRT procedures described in Chapter 3 were applied to detect DIF in the various samples. Twenty replications for each procedure were conducted for each of the conditions.

4.2.4 Evaluation Criteria

The chi-square statistic at a .01 level of significance was used to identify items that displayed DIF. All items that were identified, DIF and non-DIF, were recorded. In addition, DIF items that were not flagged, i.e. type II errors, were also recorded.

4.3 Results and Discussion

The results of this study are reported in three sections. In the first section, the effect of sample sizes on the detection rates are presented, in the second section,

the effect of setting the mean ability distribution of GR3 to -1.0 is presented. In the third section, a brief summary of the findings is given.

4.3.1 Effect of Sample Size

The results of the PIRT procedures are reported in Table 4.2 for target groups GR1, GR2 and GR3, with sample sizes of 500, 300 and 100, respectively. The data are reported in the same way as data in Chapter 3. Since sample sizes were lower, it was expected that the overall detection rates would be lower than results reported in Chapter 3, while the detection rate in GR2, with a larger sample size, was expected to be higher than that in GR3 (Rogers, 1989; Clauser, 1993).

For PIRT 1 the results obtained were not very encouraging. When GR2 was under investigation, only 40% of the DIF items were correctly flagged, while only 3% were flagged from non-target GR3. The false positive error rate, however was only 2%. For target GR3, only 25% of the DIF items were detected, while 7% were detected from non-target GR2 and the false positives error rate was 3%. Thus, when item parameter estimates were obtained from only a single analysis (that is, the total group), at most, only two-fifths of the DIF items were detected when the target groups consisted of a sample size of 300, while only a quarter of the DIF items were detected when the sample size of the target group was small (i.e., 100).

Based on the results from Table 3.2, higher detection rates had been expected for PIRT 1. While, the low detection rates observed are primarily due to lower sample sizes, it seems likely too that combining data to increase sample sizes could

have resulted in contamination of the data, and thus erroneous (i.e. misfitting) item parameter estimates were obtained (Ellis & Kimmel, 1992; Linn & Harnisch, 1981).

The results obtained with PIRT 2 and PIRT 3 were much better than with PIRT 1, since the estimates were based on uncontaminated data. For PIRT 2, the detection rate in target GR2 was 71%, with a 9% false positive error rate, and 34% in target GR3 with a 7% false positive error rate. For PIRT 3, 85% of the DIF items were detected in target GR2, with a 6% false positive error rate, while in target GR3, the detection rate was 32% with a false positive error rate of 4%. In the non-target groups GR2 and GR3 respectively, only 8 and 11% of the items were detected.

Compared to PIRT 1, the detection rate in target GR2 was approximately one-half times more for PIRT 2 (71% compared to 40%) and more than double for PIRT 3 (85% compared to 40%). However, the type I error rate for PIRT 2 was 7% higher than that of PIRT 1, and for PIRT 3, it was 4% higher. For GR3, the detection rates across the three PIRT procedures were approximately the same, that is 25%, 34% and 32%, while the false positive errors were 3%, 7% and 4% for PIRT 1, 2 and 3, respectively. Clearly, with a small sample size, many DIF items remained undetected.

Two factors contributed to the higher detection rate for PIRT 2 and PIRT 3 in this analysis. The first was data contamination. For PIRT 2 and PIRT 3, the item parameter estimates were based on (uncontaminated) data that excluded the target group. This effect is evident if target GR2 is compared across the PIRT procedures. The second factor was sample size. Detection rates for GR3, with low

sample size, were lower than that for GR2. In PIRT 1, a combination of both factors resulted in the lowest detection rate for GR3. One possible reason for the low detection rate in GR3 is that the residuals were calculated using a sample size of only 100. The number of examinees assigned to each of the ten score categories was probably too low to provide sufficient power for detecting the size of DIF which was simulated. There is also a possibility that the test statistic is not distributed as a chi-square statistic with the smaller sample sizes. However, the point is that the detection rate was low.

When GR1 was the target group, 8% and 3% of the DIF items in non-target GR2 and GR3, respectively were detected, while for PIRT 3, the detection rate was 68 and 28%, respectively. The false positive error rate for PIRT 1 was 2% compared to 15% for PIRT 3. The low detection rate in GR1 for PIRT 1 was expected since GR1 was compared to the total combined sample, of which 55% of the data were responses from GR1. The responses from GR2 only represented 33% of the data, while for GR3, it was 11%. With the item parameter estimates based on this data, it was difficult for any of the items to be detected from GR2 and GR3. For PIRT 3, item parameter estimates were based on uncontaminated data, as GR1 was compared to GR2 and GR3 combined. Although the sample size was smaller (than that in PIRT 1) the contribution from GR2 was 75% and 25% from GR3. Thus, as expected, the detection rate was higher in target GR2 than GR3. However, the false positive error rate was relatively high as well.

Table 4.3 presents the results obtained for the LR procedures. As expected, the detection rate was higher in target GR2 than in GR3: 70 and 34% of the DIF

items were correctly flagged for target GR2 and GR3, respectively, while the false positive error rate was 1% in both groups. These results are approximately the same as PIRT 2, although the false positive error rate is much lower. When GR2 was compared to GR3 (that is 20% DIF between these two groups), 37% (4.44 items) of the DIF items were detected. Of this, 42% (2.52 items) were detected in GR2 and 32% (1.92 items) in GR3.

For LR 2, the detection rate in both target GR2 (75%) and GR3 (37%) was higher than that found in the corresponding groups for LR 1, and the false positive error rate was the same at 1%. The fact that estimates were based on increased sample sizes could explain this slight improvement in the detection rate. In the non-target group GR2, only 1% of the DIF items were detected, while in non-target GR3, 8% of the DIF items were detected. In comparison to PIRT 3, the detection rate for LR 2 was 10% lower in GR2, and 5% higher in GR3. In the groups 23 v 1 comparison, 45% of the DIF items in GR2 were correctly flagged, while in GR3 it was 9%. Compared to target GR1 for PIRT 3, this represents a reduction in the detection rate of 23% in GR2 and 17% in GR3. However, it must be noted that the false positive error rate in for LR 2 was 14% less than for PIRT 3. Thus, the higher detection rate is understandable, since for PIRT 3 the overall number of items flagged was high.

In Tables 4.4 and 4.5, the effect of sample size on the percentage of item types detected are reported for the PIRT and LR procedures, respectively. Like Tables 3.4 and 3.5, the item number, its location, a- and b-values, the DIF sizes as well as the detection rate for each DIF item is reported for each of the target and

non-target groups. For the PIRT procedures, the detection rate pattern for the different item types was similar to that observed in the previous chapter (see Table 3.4). That is, for uniform DIF items, the detection rate increased as the DIF size increased, while for non-uniform DIF items, items with lower b-values were more easily detected. However, the reduction in sample sizes greatly decreased the number of items detected, especially in GR3.

For PIRT 1, 20% of the items with DIF size of .4, 55% of items with DIF size of .6 and 95% of items with DIF size of .8 were detected for uniform DIF items in target GR2. For the non-uniform DIF items, approximately one third of the moderate and low difficulty items and 15% of the high difficulty items were detected. In target GR3, the detection rate was slightly lower for the uniform DIF items, that is 15, 40 and 95% of items with DIF size of .4, .6 and .8 respectively were correctly flagged. No non-uniform DIF items were detected. When GR1 was the target group, the detection rate for the different item types was very random. In GR2, only 5% of the uniform DIF items with DIF size of .4, and 20% with DIF size of .8 were detected, while for the non-uniform DIF items, 5% of the moderate difficulty and 15% of the low difficulty items were detected. In GR3, none of the uniform DIF items were detected, while 5% of the moderate difficulty and 10% of the low difficulty non-uniform DIF items were detected.

For PIRT 2, the detection rates for the uniform DIF items were relatively high in target GR2, that is, 70, 85 and 100% for items with DIF sizes of .4, .6 and .8, respectively. For the non-uniform DIF items, the detection rate was approximately 50% for all items. In target GR3, the detection rate for uniform DIF

items with DIF sizes of .4, .6 and .8 was 30, 55 and 95%, respectively. For the non-uniform DIF items, the detection rate was 10% for the moderate and high difficulty items, and 0% for the low difficulty items.

For PIRT 3, over 75% of all DIF items in GR2 were accurately flagged, while in non-target GR3, the detection rate for all items was less than 20%. In target GR3, 20, 55 and 75% of the uniform DIF items with DIF sizes of .4, .6 and .8, respectively, were flagged. For the non-uniform DIF items, 10% of the moderate difficulty, 25% of the low difficulty and 5% of the high difficulty items were accurately flagged. For non-target GR2, the detection rate for all items was also less than 20%.

The results indicate that uniform DIF items with high DIF sizes were easier to detect than items with low and moderate DIF sizes, in the respective target groups for all the PIRT procedures, even when the sample size was as low as 100. In addition, the detection rate for the uniform DIF items was higher than that for the non-uniform DIF items. For the non-uniform DIF items, items with low difficulty were easier to detect than items with moderate difficulty, while items with high difficulty were the hardest to detect. Of the three PIRT procedures, the highest detection rate was when estimates were not contaminated and the sample sizes of the target groups were relatively large, that is in GR2 for PIRT 3. The PIRT 1 did not appear to be able to detect non-uniform DIF items well, and only seemed to detect a significant amount of uniform DIF items with large DIF sizes.

The effect of sample sizes on the percentage of item types detected using LR procedures are reported in Table 4.5. In all instances, the detection rate pattern was

similar to that of Table 3.4 (chapter 3) and 4.4. Better results were obtained for LR 2 than for LR 1, although the differences were only of the magnitude of 5 or 10% between the respective target groups. Compared to the PIRT procedures, the detection rate was slightly better for the PIRT procedures in target GR2 as well as non-target groups 2 and 3, including when GR1 was under investigation. In target GR3, the detection rate were approximately the same for all items.

4.3.2 Effect of Different Mean Ability Distribution

Table 4.6 shows the results when the mean ability distribution for GR3 was set to one standard deviation lower than that of GR1 and GR2. The sample sizes for GR1, GR2 and GR3 were kept the same (i.e., 500, 300 and 100, respectively). Slightly lower detection rates in all groups were expected, as compared to results in Table 4.2 and 4.3, since the effect of comparing groups with different ability distributions reduces the number of examinees available with the same abilities (Camilli & Shepard, 1994). The difference in results is especially pronounced for PIRT 1: 31 and 12% of the DIF items were correctly flagged, with false positive error rates of 2% in target GR2 and GR3, respectively. Compared to Table 4.2, this represents a 25% reduction in the detection rate in GR2 and 50% reduction in GR3. The detection rates in the non-target GR2 and GR3 was 1 and 4%, respectively.

An explanation for this low detection rate is that when target groups were combined to form the total (combined) sample, the contribution of the target groups to the data was substantially reduced, and thus being able to detect any effect in the

target groups was substantially reduced. This effect is more pronounced in GR3 as the sample size was much smaller. In addition, since the ability distributions differed, when groups are combined, the mean ability distribution of the combined sample is reduced, making it much harder to detect any DIF items in GR3 since GR3 is part of the total sample as well.

As expected, the detection rate and false positive error rate in GR2 for PIRT 2 did not change significantly, since the estimates were not affected by the change in the mean ability distribution of GR3. Seventy eight percent of the DIF items were correctly flagged, while 10% of the items were incorrectly flagged as DIF.

However, significant changes were obtained in GR3. The detection rate (93%) increased three fold, while the false positive error rate was 59%, an increase of approximately eight fold. The effect of lower sample sizes, coupled with the lower ability distributions is that not only were fewer examinees assigned to the ability score categories, but the score categories do not share a great deal of overlap. The chi-square statistic could only be only computed over a limited region of the ability continuum, and thus a greater number of items are flagged as DIF due to the lack of adequate data. The point is that it is not known how the chi-square statistic (Yen's Q1) functions in small samples, and therefore some of the problems with the PIRT results may be due to this factor.

Dramatic changes in both the detection rate and false positive error rate were also noted in PIRT 3, as expected, since estimates were obtained on combined data and thus the mean ability score for GR3 affected all estimates. The detection rate in target GR2 was 71% (14% decrease), in non-target GR3 it was 25% (14% increase)

and the false positive error rate was 11% (5% increase). This result is very similar to that obtained in PIRT 2. In target GR3, the detection rate was 91% (approximately 3 fold increase), 61% in non-target GR2 (7 fold increase), and the false positive error rate was 43% (10 fold increase). The results is similar to those obtained for PIRT 2. That is, a significant number of items were flagged as DIF. Therefore the detection rate as well as the false positive error rate was unusually high.

Table 4.7 presents the results obtained for the LR procedures when the mean ability distribution for GR3 was set to -1.0. For LR 1, 75% of the items were detected for target GR2, while for target GR3, 28% of the items were detected. However, the false positive error rate for GR2 was 1%, and 2% for GR3. For LR 2, the results obtained were exactly the same as that obtained in Table 4.3 in target GR2, with a detection rate of 75%, 1% for non-target GR3 and 1% false positive errors. When GR3 was investigated, the results showed a slight decrease, that is, 34% of the DIF items were correctly flagged in target GR3 (3% decrease), 3% in non-target GR2 (5% decrease) and 0% false positive errors (1% decrease). It appears that the change in mean ability distribution for GR3 did not dramatically affect the LR results. This is probably because the LR procedure, as used in this study, did not take underlying ability distributions into account when computing differences between groups on an item.

In Tables 4.8 and 4.9, the detection rates are reported for the different item types when the mean ability distribution of GR3 was set to one standard deviation lower than that of the other groups. For the PIRT 1, the detection rates for the

uniform DIF items increased with increased DIF sizes in target GR2, while for the non-uniform DIF items, the detection rate was highest for the low difficulty items (20%), and lowest for the high difficulty items (10%). In target GR3, 5% of the uniform DIF items with DIF sizes of .4 and .6 were detected while 50% of DIF items with DIF size of .8 were detected. Of the non-uniform DIF items, only 5% of the low difficulty items were detected. In non-target GR2 and GR3, hardly any items were detected besides the uniform DIF items with a high DIF size (5 and 15% respectively). When GR1 was under investigation, the detection rate was either 5% or 0 for all items, besides in GR2, where 35% of the items with uniform DIF size of .8 were flagged.

For PIRT 2, the detection rate was approximately the same as that obtained in Table 4.6 in target GR2, since in both cases, the data were not affected by the lower mean ability distributions of GR3. However, in target GR3, 100% of all the uniform DIF items, and over 80% of non-uniform DIF items were detected. While this result may appear to be desirable, it must be noted that the false positive errors were extremely high, that is 59%.

For PIRT 3, the detection rates for item types were expected to be the lower than that reported Table 4.2, where the mean ability distribution of all groups was 1. However, the same pattern, that is, increasing detection rates with increasing DIF size for uniform DIF items, and higher detection rates for low difficulty non-uniform DIF items were expected. In target GR2, the detection rate pattern observed for the uniform DIF items was as expected, while for the non-uniform DIF items, a slight increase in the detection rate was noted for the DIF items with

moderate and low difficulty. However, huge increases in detection rates were observed for all items that were flagged in GR3, whether GR3, GR2 or GR1 was under investigation. This can be attributed to the high percentage of false positive errors.

For the LR procedures, the change in the mean ability distribution of GR3 did not seem to result in significant changes, although the detection rates were lower. For LR 1, the percentage of items flagged in target GR2 were approximately the same as that in Table 4.5, In target GR3, the detection rate were much lower, differing by between 5 and 15%, especially for the non-uniform DIF items. Similar differences were observed when GR2 was compared to GR3. For LR 2, detection rates in target GR2 were 10% lower for the low and medium size uniform DIF items, while detection rates for the low difficulty non-uniform items were the same and showed an increase of 15% for the moderate, and 5% for the high difficulty items. Only 5% of the items were detected in non-target GR3. In target GR3, approximately one-third of the uniform DIF items with low and medium DIF sizes were detected, while the detection rate for the high DIF size items was 95%. For the non-uniform DIF items, the detection rate was less than 15%. In non-target GR2, no uniform DIF items were flagged, while only 5% of the low difficulty non-uniform DIF items were flagged. When GR1 was compared to GR2 and GR3 combined, the detection rate in GR2 was 35, 70 and 100% for the uniform DIF items respectively, while for the non-uniform DIF items, it was 30, 80 and 10% for the moderate, low and high difficulty items. This is slightly higher than the

percentages reported in Table 4.5. In GR3, only 5% of the uniform DIF items with large DIF sizes were flagged.

4.3.3 Summary

It is well documented in the DIF literature that small sample sizes and lower ability distributions in the focal group result in a lower detection rate for DIF items (Budgell, 1992; Camilli & Shepard, 1994; Clauser, 1993; Hambleton & Rogers, 1989; Rogers, 1989; Shepard, Camilli & Williams, 1985; Swaminathan & Rogers, 1990). The results of this study confirmed this finding. Across both the PIRT and LR procedures, detection rates were higher for GR2 (sample size of 300) than for GR3 (sample size 100), but lower than the results obtained in Chapter 3, where sample size for each group was 1000. When the mean ability of one group was reduced, the detection rate for the PIRT procedures decreased slightly, while for the LR procedures, it was approximately the same.

For the PIRT procedures, the results obtained for PIRT 1 when sample sizes differed were disappointing. Only 40% of the DIF items were detected when the sample was 300 and only 25% of the DIF items were detected when the sample was 100. However, for PIRT 2, 71 and 34% of the DIF items were correctly flagged in GR2 and GR3, respectively, with a false positive error rate less than 10%. For PIRT 3, 85 and 32% of the DIF items were correctly flagged in target GR2 and GR3, respectively, and the false positive error rate was less than 6%. When sample sizes vary and all groups are relatively small in size, PIRT 3 is clearly the procedure of choice of the pseudo-IRT procedures.

When the ability distribution of GR3 was set to -1.0 (Table 4.6), the greatest effect noted was in GR3, when compared to results obtained when the ability distributions were equal (Table 4.2). For PIRT 1, the detection rate decreased by 25% in GR2 (sample size of 300) and by 50% in GR3 (sample size 100). While the detection rate in GR2 for PIRT 2 did not change significantly, in GR3 the detection rate increased three fold and the false positive error rate increased eight fold. For PIRT 3, the detection rate in GR2 decreased by 14%, but the false positive error rate increased by 5%, while in GR3, the detection rate increased three fold and the false positive error rate increased ten fold.

In both target GR2 and GR3, the detection and false positive error rates obtained were relatively similar for PIRT 2 and PIRT 3. In GR2, this result is probably due to the moderate sample size of GR2 and small sample size of GR3. The effect size (contribution) of GR3 when GR2 was compared to GR1 and GR3 combined was probably substantially reduced, and thus the comparison was more like the GR2 v GR1 comparison used in PIRT 2. In GR3, a high detection rate was obtained for both PIRT 2 and PIRT 3 since a significant number of items were flagged as DIF, which included those items that were simulated as DIF. This result was probably due to the fact that the test statistic used (i.e., chi-square statistic) is not stable when sample sizes are small. Thus, significant differences in performance were noted on most items, and consequently, a high number of the items were flagged as DIF. When sample sizes were small and ability distributions differed, PIRT 3 was the procedure of choice when pseudo-IRT procedures were used to compare multiple groups, since detection rates for target GR2 and GR3 were

relatively high. However, the high false positive error rate is problematic. A better alternative would be to use the LR procedures.

For the LR procedures, the change in the ability distribution of GR3 did not seem to have any significant impact on the results. This could be attributed to the fact that LR procedures are model based, and thus do not take the underlying ability distributions into account when comparing the performance of two groups on an item. Instead, for the LR procedures, the slopes and intercepts for the two groups on an item are compared. Between the two LR procedures, LR 2 is preferred as the detection rate in both target GR2 and GR3 was higher. However, detection rates only differed by approximately 6%, while the false positive error rate only differed by 1% in target GR2 and GR3. Compared to the PIRT procedures, the LR procedures had a similar, if not higher detection rate in target GR2 and GR3, but the false positive error rate was significantly lower.

Compared to the PIRT procedures, the LR procedures are certainly the procedure of choice. The better results obtained with the LR procedures is probably because of: (1) greater stability (i.e., less sampling errors) in estimating the parameters, and (2) the malfunctioning of the Q1 statistic (Yen, 1981) with small samples. While the detection rate in GR3 was still low, at least a third of the DIF items were correctly detected. This is certainly better than doing no analysis at all.

Table 4.1

Group Sample Sizes, Mean Abilities and Standard Deviations

Group	Sample size	1st Analysis		2nd Analysis	
		Mean	SD	Mean	SD
1	500	0	1	0	1
2	300	0	1	0	1
3	100	0	1	-1	1

Table 4.2

Effect of Sample Size on the Percentage of DIF Items (over 20 replications)
 Detected Using the for Pseudo-IRT Procedure 1, 2 and 3
 (Sample size: GR 1 = 500, GR 2 = 300, GR 3 = 100)

Target group	PIRT 1 (TOT)		PIRT 2 (FR)		PIRT 3 (ADJ)	
	Detection Rate DIF items in GR 2	False positives GR 3	Detection Rate	False positives	Detection Rate DIF items in GR 2	False positives GR 3
GR 1	8	3	N/A	1	68	15
GR 2	40	3	71	9	85	6
GR 3	7	25	34	7	8	4

Table 4.3

Effect of Sample Size on the Percentage of DIF Items (over 20 replications)
 Detected Using the LR Procedure 1 and 2
 (Sample size: GR 1 = 500, GR 2 = 300, GR 3 = 100)

Comparisons	LR 1		LR 2		False positives	Comparison	LR 2		False positives
	Detection rate DIF items in GR 2	Detection rate in GR 3	DIF items in GR 2	Detection rate in GR 3			DIF items in GR 2	Detection rate in GR 3	
Groups 1 v 2	70	--	75	1	1	Groups 13 v 2	75	1	1
Groups 1 v 3	--	34	8	1	1	Groups 12 v 3	8	37	1
Groups 2 v 3	42	32	45	1	1	Groups 23 v 1	45	9	1

Table 4.4

Effect of Sample Size on the Percentage of DIF Items Detected Using Pseudo-IRT Procedures

Item	Origin	DIF ITEM		PIRT 1			PIRT 2			PIRT 3				
		a	b	a and b Values	Size	GR 1	GR 2	GR 3	GR 1	GR 2	GR 3	GR 1	GR 2	GR 3
1	2	1	-2	.2	.4	5	20	0	-	70	-	55	80	0
2	2	1	-.3	.3	.6	0	55	0	-	85	-	85	95	5
3	2	1	-.4	.4	.8	20	95	10	-	100	-	100	100	15
		<u>b</u>	<u>a1</u>	<u>a2</u>										
4	2	0	.6	1.4	.8	5	30	10	-	50	-	50	75	5
5	2	-1	.6	1.4	.8	15	25	15	-	55	-	75	90	20
6	2	1	.6	1.4	.8	0	15	5	-	50	-	40	70	5
		<u>a</u>	<u>b1</u>	<u>b2</u>										
55	3	1	-.2	.2	.4	0	0	15	-	-	30	30	10	20
56	3	1	-.3	.3	.6	0	0	40	-	-	55	30	20	55
57	3	1	-.4	.4	.8	0	0	95	-	-	95	55	5	75
		<u>b</u>	<u>a1</u>	<u>a2</u>										
58	3	0	.6	1.4	.8	5	0	0	-	-	10	20	5	10
59	3	-1	.6	1.4	.8	10	15	0	-	-	5	15	15	25
60	3	1	.6	1.4	.8	0	0	0	-	-	10	15	10	5

Table 4.5

Effect of Sample Size on the Percentage of DIF Items Detected Using Logistic Regression Procedures

Item	Origin	DIF ITEM		Size	LR 1			LR 2		
		a and b Values	a		1 v 2	1 v 3	2 v 3	13 v 2	12 v 3	23 v 1
1	2	a	1	.4	55	-	15	65	5	30
2	2	b	1	.6	80	-	35	90	0	50
3	2	a	1	.8	100	-	85	100	20	90
		a1	<u>a2</u>							
4	2	a	0	.8	55	-	40	55	5	25
5	2	b	-1	.8	90	-	40	90	10	60
6	2	a	1	.8	40	-	35	50	5	15
		a1	<u>b2</u>							
55	3	a	1	.4	-	25	10	5	25	10
56	3	b	1	.6	-	45	55	0	50	15
57	3	a	1	.8	-	90	85	5	95	10
		a1	<u>a2</u>							
58	3	a	0	.8	-	15	25	0	20	5
59	3	b	-1	.8	-	20	15	0	20	10
60	3	a	1	.8	-	10	5	0	10	5

Table 4.6

Effect of Sample Size and Ability Distribution on the
 Percentage of DIF Items Detected Using Pseudo-IRT Procedures
 (Sample Size: GR 1 = 500, GR 2 = 300, GR 3 = 100; Mean GR 3 = -1)

Target group	PIRT 1 (TOT)		PIRT 2 (FR)		PIRT 3 (ADJ)			
	Detection Rate DIF items in GR 2	GR 3	False positives	Detection Rate	False positives	Detection Rate DIF items in GR 2	GR 3	False positives
GR 1	9	3	1	N/A	1	75	47	24
GR 2	31	1	2	78	10	71	25	11
GR 3	4	12	2	93	59	61	91	43

Table 4.8

Effect of Varying Ability Distributions on the Percentage of Item Types Detected Using Pseudo-IRT Procedures

Item	Origin	DIF ITEM			PIRT 1			PIRT 2			PIRT 3				
		a and b values		Size	Detection Rate			Detection Rate			Detection Rate				
		a	b1		b2	GR 1	GR 2	GR 3	GR 1	GR 2	GR 3	GR 1	GR 2	GR 3	
1	2	1	-2	.2	.4	0	10	0	0	0	75	-	95	25	70
2	2	1	-.3	.3	.6	5	40	10	10	95	-	-	100	55	60
3	2	1	-.4	.4	.8	35	90	0	0	100	-	-	100	95	70
		<u>b</u>	<u>a1</u>	<u>a2</u>											
4	2	0	.6	1.4	.8	5	15	0	0	65	-	-	45	85	70
5	2	-1	.6	1.4	.8	5	20	0	0	75	-	-	55	100	90
6	2	1	.6	1.4	.8	5	10	15	15	60	-	-	55	65	5
		<u>a</u>	<u>b1</u>	<u>b2</u>											
55	3	1	-.2	.2	.4	5	0	5	5	-	-	100	65	15	100
56	3	1	-.3	.3	.6	0	0	5	5	-	-	100	75	30	100
57	3	1	-.4	.4	.8	5	5	50	50	-	-	100	90	50	100
		<u>b</u>	<u>a1</u>	<u>a2</u>											
58	3	0	.6	1.4	.8	0	0	0	0	-	-	95	20	25	90
59	3	-1	.6	1.4	.8	5	0	5	5	-	-	85	20	15	85
60	3	1	.6	1.4	.8	0	0	0	0	-	-	80	10	15	80

Table 4.9

Effect of Varying Ability Distributions on the Percentage of Item Types Detected Using Logistic Regression Procedures

Item	Origin	DIF ITEM		a and b values	Size	LR 1						LR 2		
		a	b			1 v 2	1 v 3	2 v 3	13 v 2	12 v 3	23 v 1	13 v 2	12 v 3	23 v 1
1	2	1	1	-.2	.2	.4	60	-	0	55	0	35		
2	2	1	1	-.3	.3	.6	85	-	30	80	0	70		
3	2	1	1	-.4	.4	.8	100	-	45	100	0	100		
		<u>b</u>		<u>a1</u>	<u>a2</u>									
4	2	0	0	.6	1.4	.8	70	-	30	70	5	30		
5	2	-1	0	.6	1.4	.8	95	-	30	90	0	80		
6	2	1	0	.6	1.4	.8	40	-	45	55	10	10		
		<u>a</u>		<u>b1</u>	<u>b2</u>									
55	3	1	1	-.2	.2	.4	-	15	20	0	25	0		
56	3	1	1	-.3	.3	.6	-	35	35	0	30	0		
57	3	1	1	-.4	.4	.8	-	90	80	0	95	5		
		<u>b</u>		<u>a1</u>	<u>a2</u>									
58	3	0	0	.6	1.4	.8	-	0	10	5	10	0		
59	3	-1	0	.6	1.4	.8	-	20	10	0	15	0		
60	3	1	0	.6	1.4	.8	-	5	10	0	10	0		

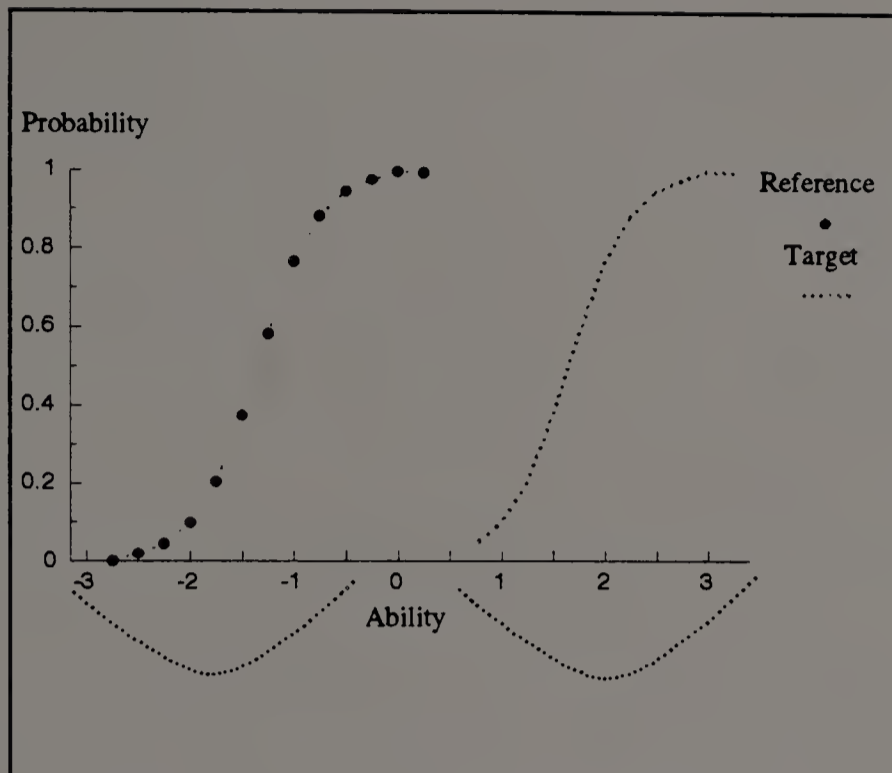


Figure 4.1

An example showing two ICCs that are far apart but when their associated ability distributions are considered, the differences in ICCs affects no one.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Summary

At present, small sample sizes are a problem with many procedures used for detecting differential item functioning (DIF). DIF is said to be present when examinees of the same ability but belonging to different groups have different probabilities of success on an item. However, most procedures currently used focus on the detection of DIF using two-group and/or pairwise comparisons only, even when multiple groups are compared, for example in cross-cultural/national studies. A consideration of all the possible combinations is a possibility but such an approach has several drawbacks including the inflation of Type I errors and considerable expansion of the amount of work needed.

In this study, the performance of two techniques to address the issue of small sample size by simultaneously detecting DIF in more than two groups was investigated. Specifically, the performance of the pseudo-IRT (PIRT) method proposed by Linn and Harnisch (1981) was compared to the logistic regression (LR) procedure proposed by Swaminathan and Rogers (1990). Both procedures: (1) can be extended to simultaneously detect DIF in multiple groups, (2) give similar results to the widely accepted ICC method (for large samples). Their performance with small samples is less well known, though the PIRT procedure has been recommended for use with small samples, and there are some results to suggest LR is successful in detecting DIF with samples as small as 250, and (3) appear to produce stable estimates when used with small samples.

Two separate studies were conducted to investigate (1) the viability of using PIRT and LR procedures, and (2) the effect of sample size and ability distribution on the DIF statistic. In the first study, the sample sizes of three groups were set at 1000 examinees each, and ability distributions were chosen to be equal, with mean 0 and standard deviation 1. For the PIRT procedure, three different estimation procedures were used to obtain the item and ability parameters. For procedure 1 (PIRT 1), the parameters were estimated only once using the entire sample of examinees. In procedure 2 (PIRT 2), the item parameter estimates were obtained from group 1, and then held fixed to obtain the ability parameter estimates for groups 2 and 3. In procedure 3 (PIRT 3), the item parameter estimates were obtained for each of the different combined samples that excluded examinees from the target group under investigation (i.e. adjusted group). These item parameter estimates were then held fixed to obtain ability parameter estimates for the respective target groups.

For each procedure, members of each target group were divided into 10 categories based on their ability scores so that there were equal number of people in each group. Within each score category, the difference between the observed and estimated proportion correct for every item for each target group was computed. Items with a chi-square statistic significant at the .01 level were classified as exhibiting DIF.

For the LR procedure, two different methods for estimating the parameters and comparing the groups were conducted. In procedure 1 (LR 1), each group was compared to each other, that is, group 1 vs 2; group 1 v 3; and group 2 vs 3; while

in procedure 2 (LR 2), each group was compared to its respective adjusted group, that is groups 1 and 2 vs 3; groups 1 and 3 vs 2; groups 2 and 3 vs 1. The former procedure is comparable to PIRT 2, while the latter procedure is comparable to PIRT 3.

The study was conducted on data simulated to fit a unidimensional three parameter IRT model, using the computer program DATAGEN (Hambleton & Rovinelli, 1973). Data were simulated for three groups only as: (1) techniques that are applicable to three groups can readily be adapted for use with more groups, and (2) the application of statistical techniques and analysis of data was simplified.

To simulate DIF, item differences were quantified in terms of the area between the curves for any two groups. The same type and amount of DIF, 10%, was simulated in groups 2 and 3 using group 1 as a reference (no DIF). Items were simulated to exhibit both moderate to high DIF, as well as uniform and non-uniform DIF. A test length of sixty items was used as this is both within the range of typical standardized test lengths, and allowed for a reasonable number of items to be studied for DIF. Twenty replications for both the PIRT and LR procedures were conducted. All items flagged, DIF and non-DIF, as well as DIF items that were not flagged were recorded.

In the second study, the two conditions investigated were the effect of sample size and varying ability distributions on the PIRT and LR procedures. A minimum size of 100 was selected because this is typically found in practice, and samples below 100 respondents are unlikely to provide stable item parameter estimates. In addition, the ability distributions were varied as in practice it is more likely for the

different cultural/ethnic groups to have lower mean abilities than the majority group. In the first analysis, sample sizes for groups 1, 2 and 3 were set at 500, 300 and 100, respectively, while the ability distributions were held fixed at mean 0 and standard deviation 1. In the second analysis, the same sized samples were used, except that the mean ability distribution for group 3 was set at -1.0.

The results of the first simulation study indicated that all three PIRT, and both LR procedures were viable techniques for detecting DIF in multiple group comparisons. The PIRT procedures were able to detect over 94% of the DIF items simulated. The false positive error, however, was low for PIRT 1, high for PIRT 2, and very high for PIRT 3. Both LR 1 and LR 2 had similar high detection rates, but significantly lower false positive errors. PIRT 1 proved to be the procedure of choice since only a single estimation was required, a significant amount of DIF was accurately detected, and the false positive errors were relatively low.

In the second study, when the sample sizes of the groups were smaller, the percentage of DIF items detected decreased considerably for all estimation procedures and conditions. With PIRT 1, where estimates were contaminated because data from different groups were combined, less than 50% of the DIF items were detected. However, with PIRT 2 and PIRT 3 (uncontaminated, independent estimates), over 70% of the DIF items were detected in target GR2 (sample size 300), while in target GR3 (sample size 100), the detection rate was about 40%. The false positive errors in both groups was about 10%. Both LR 1 and LR 2 (see Table 4.2) produced similar results in target GR2 (75% of the DIF items were detected),

with 1% false positive errors. However, in target GR3, only about 34% of the DIF items were detected.

When the ability distribution of group 3 was set at -1.0, PIRT 2 showed the highest detection rate in target GR2. In target GR3, the detection rate for PIRT 2 and PIRT 3 was relatively high but the false positive error rate was high as well. Generally, the LR procedures had better detection rates and a lower false positive error rate. In target GR2, both LR 1 and LR 2 detected approximately two-thirds of the DIF items, while in target GR3, the detection rate was about a third of the DIF items.

Additional analysis was also conducted for the specific type of items detected in each group with the different estimation procedures. Generally, uniform DIF items with higher DIF sizes were easier to detect, even when sample sizes were small. The LR procedures had a higher detection rate for the non-uniform DIF items as compared to the PIRT procedures. Also, items detected across the respective PIRT and LR procedures were consistent in both studies.

PIRT and LR procedures are certainly viable options for simultaneously detecting DIF in multiple groups. When sample sizes are large, PIRT 1 appears to be the procedure of choice. When sample sizes are small (i.e. as small as 100), the procedure of choice appears to be LR 2, in both situations when the mean ability distributions are equal and when they are not.

5.2 Significance of the Findings

The use of simultaneous DIF detection techniques in multiple groups enable researchers to obtain information from those groups that consist of small sample sizes, instead of merely excluding these groups from any analysis. However, the information obtained, and how it is used, will depend primarily on the purpose of conducting the (DIF) study. That is, the definition of what constitutes DIF is a function of the purpose of the study. Zieky (1993) notes that even though the same methods and techniques are used, there are two main reasons that DIF studies are conducted. First, to develop test instruments that are equivalent and fair to all groups to whom these instruments are expected to be administered. Second, to ensure that scores from test instruments are not biased in favor of or against any group or groups that have been administered these instruments, because it is generally recognized that even the use of the most rigorous process cannot guarantee complete DIF-free test instruments. In practice, the former reason dictates the use of pre-testing while the latter requires the use of post-hoc analysis. The reasons for conducting DIF studies and its implications are further discussed in the next section.

As noted earlier, when sample sizes are large and ability distributions are approximately equal, PIRT 1 appears to be the procedures of choice (of the procedures studied). The advantage of using PIRT 1 is that it is easy to implement and is not time consuming to use since only a single estimation is required. However, the use of PIRT 1 may not be feasible in practice since other, more reliable DIF detection techniques that are applicable in situations where large sample sizes are available may be preferable. For example, Ellis and Kimmel (1992) used

the IRT procedure (Lord's Chi-square statistic) to compare three groups with samples of 200 and more. In addition, when sample sizes in the target group were small, PIRT 1 was only able to detect a small percentage of the DIF items.

When the sample sizes were smaller, LR 2 was the preferred procedure, whether mean ability distributions were equal or unequal, as detection rates were relatively higher and the false positive error rates significantly lower than the PIRT procedures. However, it must be noted that this result could be because of the malfunctioning of the Q1 statistic (Yen, 1981) with small samples. The advantage of LR 2 is that it is easy to implement, it is readily available in most statistical packages, and can readily be adapted to include covariates so that possible reasons for DIF can be researched. However, since only parameters were compared in the logistic regression procedure, the underlying ability distributions were not taken into account when the performance of examinees on an item are compared. Thus an item may exhibit DIF in theory (see Figure 4.1), but in practice, this DIF has limited influence on examinees (Wainer, 1993b). The point is that while higher detection rates can be expected when LR procedures are used, these detection rates may not necessarily provide researchers with any meaningful information regarding DIF.

As argued by Wainer (1993b), model-based procedures that do not account for underlying ability distributions (i.e., the LR procedure as applied in this study) are not necessarily preferable to weighted procedures, that is, procedures that incorporate information about the score distributions, for example the PIRT procedure, in the DIF detection process. However, the disadvantage of weighted

procedures is that they are sample dependent. That is, the number of people in each ability interval influence considerably whether an item is flagged as DIF. For different sample sizes, the number of persons in any ability interval is expected to differ, which has consequences on whether differences in performance between groups on an item can be detected.

5.3 Implications and Recommendations

The implications for assessment practitioners is that when multiple groups are to be compared and sample sizes are small, both the PIRT and LR techniques could be used to detect DIF. Since an analysis with some validity is much better than no analysis at all, it is recommended that groups with samples as low as 100 be included in the analysis, even though only a fraction of the DIF items will be detected. In addition, when ability differences of the groups being compared differ, practitioners need to seriously consider the advantages and disadvantages of weighted procedures, that are more likely to detect 'practical DIF', versus model based procedures, that are more likely to detect 'theoretical DIF'. Alternatively, model-based procedures can be used but then the results should be considered along with the actual reference and target group score distributions.

Depending on the reason for conducting the DIF study, the two-group comparison definition of DIF may or may not be valid. For example, if instruments are developed for different language groups and are translated from a single base language, then two-group comparison procedures (that is, PIRT 2 or LR 2) are appropriate. However, in the context of post-hoc analysis, the main purpose is to

ensure that scores on an instrument for three or more groups are equivalent. Given this, the definition of DIF must take examinee responses of all groups into account since the domain of interest is on the performance of the different target groups with respect to the adjusted group (i.e., total - target), and not with respect to each other. Thus, for any comparison, only those items that are flagged when all examinees are taken into account should be regarded as DIF.

When DIF studies on multiple groups are conducted for test development purposes, studies need to be designed such that researchers are able to conduct good DIF studies. Therefore, researchers must take the responsibility to ensure that samples collected meet the minimum (at least) sizes required. To this end, prior information of groups being compared is essential as sampling designs can be developed to maximize samples collected from minority populations. For example, if Native American Indians are one of the groups of interest, researchers need to design their study to sample in those areas where there is likely to be greater participation of Native American Indians.

For post-hoc DIF analysis, researchers need to consider not only the sample size and underlying ability distributions of groups, but also the number of score categories used. When the ability distributions are modestly different it is possible for examinees who are matched on the criterion variable, to have real differences in ability. The result is an invalid matching criterion allowing impact to be misinterpreted as DIF (Clauser, Mazor and Hambleton, 1994). This effect is most pronounced at the extreme ends of the ability distributions, where the score category widths are expected to be wider. A possible solution noted by Clauser et al. (1994)

is maximize the number of score categories used (that is, score category intervals should be smaller) when ability differences exists between groups compared.

Even when sample sizes are relatively small, it is certainly much better to *some analyses* than no analyses. This recommendation seems reasonable as long as the procedure holds some validity in the situation where it is used. For example, using the pseudo-IRT procedure where the total sample size is small would probably have no validity at all because item parameter estimates would be vary unstable. The only harm that could be done, in general, is that some items will be misclassified as DIF while other items with real DIF will be misclassified as having no DIF. Even though some real DIF items will be missed, these items would certainly have been missed if no analyses were done.

Inevitably the question arises: when are sample sizes too small to render any analysis as meaningless? That is, how do practitioners recognize that they do not have enough information and thus cannot do any reliable and meaningful analysis. Zieky (1993) notes that this question must be addressed in the context of why DIF studies are conducted. At the Educational Testing Service, for example, sample sizes of at least 100 are required if DIF statistics are to be used in test development, while at least 200 people are required when DIF statistics are computed after tests are administered but before scores are reported (Zieky, 1993). In general, however, sample sizes that are less than 100 people are not used in DIF studies.

A final point regarding the use of simultaneous DIF detection procedures used in this study is that these procedures are sample dependent since, in post-hoc analyses, what constitute DIF depends on the groups that are being compared. For

example, an assessment instrument developed for examinees from Canada, Eritria, Nigeria, Palestine and South Africa, is administered to Canadians, Eriterians, and Palestinians only, the items that are detected as DIF may or may not be the same if the instrument is administered to Nigerains, Palestinians and South Africans.

5.4 Shortcomings

Several factors that could have influenced the results of this study need to be noted. First, the two-stage procedure proposed by Holland and Thayer (1988) for detecting DIF was not applied in this study. In this procedure, DIF items are identified in the first stage, and in the second stage, these items are removed from the analysis, and estimates are recalculated using the "purified" score as the criterion. The consequence of omitting the two-stage procedure to clean-up the criterion is that the accuracy of the DIF detection techniques, whatever the technique, is reduced (Clauser, 1993). Therefore, the power of both the PIRT and LR techniques to accurately detect DIF items was also reduced. Clauser (1993) notes that when test instruments contain substantial levels of DIF, the advantage of using the two-stage procedure is greater. This is especially relevant for multiple group comparisons where data are combined to increase sample sizes, since aggregating data with unknown levels of DIF only increases this number of items detected as DIF.

Second, the number of score group categories used to determine the residuals for the PIRT procedure was fixed at 10 (Yen, 1981). It is likely that if the number

of score categories used was greater, the detection rates would have changed, especially when the ability distributions differed (Clauser, et al, 1994).

Third, other factors which could have had an influence on the results, for example, the percentage of DIF simulated, the length of the test, the amount of the difference in mean ability distributions, or the group for which the mean ability distribution was reduced, were not manipulated. These and other factors might easily be manipulated in follow-up research since they are significant factors in the DIF detection process.

Last, like any other simulation study, a possible problem with this study is that the data simulated may not reflect real data. While all possible care was taken to ensure that conditions simulated were as realistic as possible, the extent to which this study reflects actual practice is unknown.

5.5 Directions for Further Research

First, these methods need to be tested using real data. Since the amount of 'true' DIF in real data is not known, both simultaneous and two-group comparison DIF detection techniques could be used to ascertain whether the same items are detected. Second, the effect of using the two-stage procedure for detecting DIF needs to be investigated in the context of simultaneous DIF detection. Based on two-group comparisons, the application of this procedure would certainly improve detection rates since the criterion used to compare examinees from different groups would not be contaminated by the DIF items. A possible simulation study could be to compare the effect of different levels of contamination of the criterion (i.e. total

score) on the detection of DIF in multiple groups. Third, it would be interesting to ascertain how well these simultaneous DIF detection methods work when the ratio of sample sizes of the groups compared are approximately equal. For example, comparing four groups, each with sample sizes of approximately 200 each. Fourth, a study to determine the minimum total sample size as well as the minimum sample size required for the respective target groups would provide useful information to practitioners. Fifth, a study to determine the 'ideal' number of score categories required and to determine how the chi-square statistic functions when PIRT procedures are used to detect DIF in multiple groups with different sample sizes, would also be useful. Last, another interesting avenue for research is to assess whether the reasons of DIF can be studied statistically. In this respect, the use of the LR procedure is ideal since covariates can easily be included in the procedure. For example, Mazor, Kanjee and Clauser (in press) used the logistic regression procedure to account for language ability when groups were compared on two ability tests. The authors found that when language ability was included in the procedure, the number of items detected as DIF reduced significantly. However, a-priori information about examinees and the testing instruments need to be available or collected as part of a carefully designed DIF study.

APPENDIX A

PERCENTAGE OF PEOPLE AND LANGUAGES SPOKEN IN SOUTH AFRICA

PERCENTAGE OF PEOPLE AND LANGUAGES SPOKEN IN SOUTH AFRICA¹

<u>Language</u>	<u>Total</u>	<u>% Total</u>
Afrikaans	5750814	18.56
English	3436717	11.09
Netherlands	7929	0.03
German	3323	0.01
Greek	12859	0.04
Italian	8949	0.03
Portuguese	48705	0.16
French	4975	0.02
Hindi	5848	0.02
Tamil	4874	0.02
Telegu	762	0.00
Gujerati	8730	0.03
Urdu	4356	0.01
Chinese	4572	0.01
Xhosa	2513411	8.11
Zulu	8354470	26.96
Swazi	953918	3.08
South-Ndebele	217508	0.70
North-Ndebele	114910	0.37
Ndebele	146088	0.47
Northern-Sotho	6458638	20.84
Southern-Sotho	2240430	7.23
Sotho	263255	0.85
Tswana	1443478	4.66
Venda	114962	0.37
Shangaan	1440932	4.65
Other	323919	1.05
<u>TOTAL</u>	<u>30986920</u>	<u>100.00</u>

¹Information from the Central Statistic Service (1992). RSA: Statistics in brief. Pretoria: Central Statistic Service.

APPENDIX B

STUDENT-TEACHER AND -CLASSROOM RATIOS BY "RACE" GROUP

STUDENT-TEACHER AND -CLASSROOM RATIOS BY "RACE" GROUP

<u>"Race" Group</u>	<u>Student-Teacher</u>	<u>Student-Classroom</u>
"African"	1:40	1:44
"Asian"	1:20	1:28
"Coloured"	1:23	1:24
"White"	1:17	1:20

REFERENCES

- Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report No. 88-2). New York, NY: College Entrance Examination Board.
- Appel, S. W. (1989). Outstanding individuals do not arise from ancestrally poor stock: Racial science and the education of Black South Africans. Journal of Negro Education, 58, 544-556.
- Bam, L., & Rice, M. (1987). Testing what kind of instrument is it? Exactly. (ERIC Document Reproduction Service No. ED 285385).
- Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. Journal of Educational Measurement, 24, 41-55.
- Bleistein, C. A. (1986). Application of items response theory to the study of differential item characteristics: A review of the literature (RR-86-3). Princeton, NJ: Educational Testing Service.
- Bollwark, J. (1991). Evaluation of IRT anchor test designs in test translation studies. Unpublished doctoral dissertation, University of Massachusetts at Amherst.
- Bollwark, J. (1992, April). Using item response models in test translation studies: A look at anchor test length. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.
- Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. School Psychology International, 12, 119-132.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology, 1, 185-216.
- Brislin, R. W. (Ed.). (1976). Translation: Application and research. New York: John Wiley.
- Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), Field methods in cross-cultural psychology (pp. 137-164). Newbury Park, CA: Sage Publishers.

- Budgell, G. R. (1992). Analysis of differential item functioning in translated assessment instruments. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Inc. Montreal, Quebec.
- Bulhan, A., B. (1981). Psychological research in Africa: Genesis and function. Race and Class, 23, 25-41.
- Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased items. Newbury Park, CA: Sage Publishers.
- Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17, 417-440.
- Christie, P. (1985). The right to learn: The struggle for education in South Africa. Johannesburg: Ravan Press.
- Clauser, B. E. (1993). Factors affecting the performance of the Mantel-Haenszel procedure in identifying differential item functioning. Unpublished doctoral dissertation, University of Massachusetts at Amherst.
- Clauser, B. E., Mazor, K. M., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. Journal of Educational Measurement, 31, 67-78.
- Cole, N. S., & Moss, P. A. (1989). Bias in test use. In Linn, R. (Ed.) Educational Measurement (pp 201 - 219). New York: Macmillan Publishing Co.
- Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and Standard differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp.137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Psychological Measurement, 3, 217-233.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp.35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalent relations with external variables are the central issues. Psychological Bulletin, 95, 134-135.

- Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. Applied Psychological Measurement, 13, 77-90.
- Drasgow, F., & Hulin, C. L. (1986). Assessing the equivalence of measurement of attitudes and aptitudes across heterogeneous subpopulations (unpublished manuscript). Urbana-Champaign, IL: University of Illinois.
- Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology, 70, 662-680.
- Dubois, P. H. (1970). A history of psychological testing. Boston: Allyn and Bacon.
- Ellis, B. B. (1989). Differential item functioning: Implications for test translation. Journal of Applied Psychology, 74, 912-921.
- Ellis, B. B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. Bulletin of the International Test Commission, 18, 33-51.
- Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. Journal of Applied Psychology, 77, 177-184.
- Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translation: An investigation using item response theory. International Journal of Psychology, 24, 661-684.
- Ellis, B. B., & Weiner, S. P. (1990). A study of the gender differences in two countries: Implications for future research. In N. Bleichrodt & P.J.D. Drenth (Eds.), Contemporary issues in cross-cultural psychology. Amsterdam, Netherlands: Swets & Zeitlinger.
- Gifford, J. A. (1983). An empirical investigation of Bayesian procedures in item response models. Unpublished doctoral dissertation, University of Massachusetts Amherst.
- Gould, S. J. (1981). The mismeasure of man. New York: W. W. Norton and Company.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed; pp. 147-200). New York: Macmillan.
- Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 54-65.

- Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. Bulletin of the International Test Commission, 18, 3-32.
- Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in detection of differentially functioning test items. European Journal of Psychological Assessment, 9, 1-18.
- Hambleton, R. K., & Cook, L. L. (1983). The robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), New horizons in testing (pp. 31-49). New York: Academic Press.
- Hambleton, R. K., & Kanjee, A. (In press). Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods. In T. Husen & T.N. Postlewaite (Eds.), International Encyclopedia of Education (2nd ed.). Oxford, UK: Pergamon Press.
- Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.
- Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 17, 73-74.
- Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publishers.
- Hills, J. R. (1989) Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Hosmer, D. W. & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley
- Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. Journal of Cross-Cultural Psychology, 16, 131-152.

- Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18, 115-142.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Application of item response theory to psychological measurement. Homewood, IL: Dow-Jones-Irvin.
- Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71, 83-94.
- Ikeda, H. (1991). Stability of test-item statistics in different cultures. Bulletin of the International Test Commission, 18, 52-64.
- Ironson, G. H. (1982). Use of chi-square and latent trait approaches of detecting item bias. In R. A. Berk (Ed.), Handbook of Methods for Detecting Item Bias. Baltimore, MD: John Hopkins University Press.
- Irvine, S. H., & Carroll, W. K. (1980). Testing and assessment across cultures: Issues in methodology and theory. In H. C. Triandis & J. W. Berry (Eds.), Handbook of cross-cultural psychology (Volume 2) (pp. 181-244). Boston: Allyn and Bacon.
- Jansen, J. (1990). Curriculum in a post-apartheid dispensation. In M. Nkomo (Ed.), Pedagogy of domination: Towards a democratic education in South Africa (pp. 325-340). NJ: Africa World Press.
- Kallaway, P. (1984). An introduction to the study of education for Blacks in South Africa. In Kallaway, P. (Ed.) Apartheid and education: The education of Black South Africans. Johannesburg: Ravan Press.
- Kanjee, A. (1993a, June). Maintaining the status quo: Certification and education in South Africa. Paper presented at the 5th Conference of North American and Cuban Philosophers. Havana, Cuba.
- Kanjee, A. (1993b, June). Interview with Luis del Toto Reyes, Director Department Independente Ingresso - (Director of Admissions and Placement), Havana, Cuba.
- Keeves, J. P. (1992, April). Technical issues in the first and second IEA science studies. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

- Kim, S. -H., Cohen, A. S., & Park, T. -H. (1993, April). Detection of differential item functioning in multiple groups. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.
- King, M., & van den Berg, O. (1991) One nation, many languages. Pietermaritzburg: Centaur Publications.
- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research, 26, 457-477.
- Lapointe, A. E., Mead, N. A., & Phillips, G. W. (1989). A world of differences: An international assessment of mathematics and science (Report No. 19-CAEP-01). Princeton, NJ: Educational Testing Service.
- Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.), Educational measurement (3rd ed; pp. 1-10). New York: Macmillan Publishing Company.
- Linn, R. L. (1993). The use of differential item functioning statistic: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 329-363). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Linn, R. L., & Harnisch, D. L. (1981). Interaction between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.
- Lonner, W. J. (1990). An overview of cross-cultural testing and assessment. In R.W. Brislin (Ed.), Applied cross-cultural psychology (pp. 56-76). Newbury Park, CA: Sage Publications.
- Lonner, W. J., & Berry, J. W. (Eds.) (1986). Field methods in cross-cultural research. Newbury Park, CA: Sage Publications.
- Malaka, M. L. (1992). Apartheid education as a mechanism for political control: The role of testing in schools and the national examination system. Paper submitted in partial fulfillment of the Comprehensive Examination Requirement, University of Massachusetts Amherst.
- Mathonsi, E. N. (1988). Black matriculation results: A mechanism of social control. Johannesburg: Skotaville Publishers.

- Mayberry, P. W. (1984, April). Analysis of cross-cultural attitudinal scale translation using maximum likelihood factor analysis. Paper presented at the meeting of the American Educational Research Association, New Orleans: LA.
- Mazor, K., Clauser, B., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-452.
- Mazor, K., Kanjee, A., & Clauser, B. (in press). Using logistic regression with multiple ability estimates to detect differential item functioning. Journal of Educational Measurement.
- McCauley, D.E., & Coleberg, M. (1983). Transportability of deductive measurement across cultures. Journal of Educational Measurement, 20, 81-92.
- McKinley, R. L. & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.
- Mellenberg, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13, 127-143.
- Mislevy, R. J., & Bock, R. D. (1990). Bilog: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville IN: Scientific Software, Inc.
- Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), Educational measurement (3rd ed; pp. 447-474). New York: Macmillan Publishing Company.
- Nkomo, M. (1990). Post-apartheid education: Preliminary reflections. In M. Nkomo (Ed.), Pedagogy of domination: Towards a democratic education in South Africa (pp. 291-325). NJ: Africa World Press.
- Norris, N. (1990). Understanding educational evaluation. New York: St. Martin's Press
- Omotoso, K. (1994, January 21-27) Don't neglect the mother tongue in favor of English. The Weekly Mail and Guardian, p. 37.
- Parshall, C. G., & Kromrey, J. D. (1992). Performance of statistical indices of test item bias in small sample sizes. Paper presented at the meeting of the American Educational Research Association, San Francisco.

- Poortinga, Y. H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S.H. Irvine & J.W. Berry (Eds.), Human assessment and cross-cultural factors (pp. 237-258). New York: Plenum Press.
- Poortinga, Y. H., & Malpass, R. S. (1986). Making inferences from cross-cultural data. In W.J. Lonner & J.W. Berry (Eds.), Field methods in cross-cultural psychology (pp. 17-46). Beverly Hills, CA: Sage.
- Poortinga, Y. H., & Van de Vijver, F. J. R. (1987). Explaining cross-cultural differences: Bias analysis and beyond. Journal of Cross-Cultural Psychology, 18, 259-282.
- Poortinga, Y. H., & Van de Vijver, F. J. R. (1988). The meaning of item bias in ability tests. In S.H. Irvine & J.W. Berry (Eds.), Human abilities in cultural context (pp. 166-183). New York: Cambridge University Press.
- Poortinga, Y. H., & Van de Vijver, F. J. R. (1991). Culture-free measurement in the history of cross-cultural psychology. Bulletin of the International Test Commission, 18, 72-87.
- Popham, J. W. (1990). Modern educational measurement: A practitioner's perspective. Englewood Cliffs, New Jersey: Prentice Hall.
- Prieto, A. J. (1992). A method for translation of instruments to other languages. Adult Education Quarterly, 43, 1-14.
- Prinsloo, R. J. (1984). Test practices and the legal control of psychological tests in the Republic of South Africa. Revue de Psychologie Applique, 34, 39-53.
- Raju, S. N., Bode R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Psychological Measurement, 3, 1-13.
- Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. Unpublished doctoral dissertation, University of Massachusetts Amherst.
- Rogers, H. J., & Swaminathan, H. (1994, April) Logistic regression procedures for detecting DIF in nondichotomous item responses. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.
- Royce, J. R. (1988). The factor model as a theoretical basis for individual differences. In S.H. Irvine & J.W. Berry (Eds.), Human abilities in cultural context (pp. 147-165). New York: Cambridge University Press.

- Sax, G. (1989). Principles of educational and psychological measurement. Monterrey, California: Wadsworth Publishing Company.
- Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.
- Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. Journal of Educational Measurement, 25, 1-13.
- Schmitt, A. P., & Crone, C. R. (1991). Alternative mathematical aptitude item types: DIF issues (RR-91-42). Princeton, NJ: Educational Testing Service.
- Schmitt, A. P., Holland, P. W., & Dorans N. J. (1993). Evaluating hypothesis about differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Shepard, L. A., Camilli, G. & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 49-58.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.
- Swartz, D. (1992). Issues in the reform of the system of certification South Africa. In E. Unterhalter, H. Wolpe, & T. Botha (Eds.), Education in a future South Africa: Policy issues and transformation (pp 136-148). Trenton. NJ: Africa World Press.
- Traub, R. E. & Lam, R. (1985). Latent structure and item sampling models for testing. Annual Review of Psychology, 36, 19-48.
- Triandis, H. C. (1976). Approaches toward minimizing translation. In R. W. Brislin (Ed.), Translation: Application and research (pp 228-243). New York: John Wiley.
- Triandis, H. C., & Berry, J. W. (Eds.). (1980). Handbook of cross-cultural psychology (Volume 2). Boston: Allyn and Bacon, Inc.
- Triandis, H. C., Bontempo, R., & Hui, C. H. (1990). A method for determining cultural, demographic and personal constructs. Journal of Cross-Cultural Psychology, 21, 302-318.

- Van de Vijver, F. J. R. (1991). Group differences in structured tests. In P. L. Dunn, S. H. Irvine, & J. M. Collis (Eds.), Advances in computer-based human assessment (pp. 397-417) Dordrecht: Kluwer Academic Publishers.
- Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing (pp. 277-307). Dordrecht: Kluwer Academic Publishers.
- Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Culture-free measurement in the history of cross-cultural psychology. Bulletin of the International Test Commission, 18, 72-87.
- Van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogenous populations: When are cultural loadings undesirable? European Journal of Psychological Assessment, 8, 17-24.
- Wainer, H. (1993a). Measurement problems. Journal of Educational Measurement, 30, 1-21.
- Wainer, H. (1993b). Model based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 123-136). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Weinberg, S. (1992, February 15-21). State does the blackboard crab-walk. The Weekly Mail, p. 49.
- Whittaker, S. B. (1990, March). A critical historical perspective on psychology in Azania/South Africa. Paper presented at the Psychology and Apartheid Conference, University of the Western Cape, Cape Town.
- Wolf, R. M. (1992, April). The use of performance assessment in IEA studies over the past 30 years: Its successes and problems, and its implications for future surveys. Paper presented at the meeting of the American Educational Research Association, San Francisco.
- Wolpe, H. (1992). Education and social transformation: Problems and dilemmas. In E. Unterhalter, H. Wolpe, & T. Botha (Eds.), Education in a future South Africa: Policy issues and transformation (pp. 1-16). Trenton, NJ: Africa World Press.
- Yen, W. M. (1981) Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337-348). Hillsdale, NJ: Lawrence Erlbaum Associates.

