INVESTIGATION OF WITHIN GROUP HETEROGENEITY IN MEASUREMENT COMPARABILITY RESEARCH

by

MARÍA ELENA OLIVERI

B.A., University of British Columbia, 1996
B.Ed., University of British Columbia, 1997
M.Ed., University of British Columbia, 2003
M.A., University of British Columbia, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES
(Measurement, Evaluation and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

January 2012

© María Elena Oliveri, 2012
ABSTRACT

This dissertation investigated the broad concern of the violation of the within group homogeneity (WGH) assumption associated with methods used to investigate differential item functioning (DIF) in educational large-scale assessments (ELSAs) in two studies. The first study investigated the WGH assumption using real data from an international ELSA of reading literacy administered to examinees from four countries: Chinese Taipei, Hong Kong, Qatar and Kuwait. It examined the degree of homogeneity of the test data structure, and the consistency in identification of DIF items between a manifest versus a latent class DIF approach. Sources of incomparability underlying DIF items related to reading comprehension processes, cognitive complexity of items and purposes of reading as well as instruction and teacher-related variables were investigated. Three latent classes (LCs) were identified indicating heterogeneity in response patterns. The two approaches identified different DIF items with the latent class approach identifying a higher number of differentially functioning items with larger effect sizes. Inconclusive results were obtained when analyzing sources of manifest DIF items. Student achievement was found to be a primary factor influencing LC membership, whereas typically used manifest characteristics (e.g., country or gender) had either small or no significant influence on LC assignment. The second study used simulated data to systematically investigate the effects of heterogeneity on accuracy of DIF detection by two manifest DIF detection methods: logistic regression and Mantel-Haenszel. Greater degrees of heterogeneity resulted in significantly reduced DIF detection accuracy.

Findings from this dissertation have important implications for DIF and validity research. The first study presents evidence of within group heterogeneity in international assessments and therefore, contributes to research that suggests that the WGH assumption needs to be tested for meaningful manifest DIF analysis. Moreover, findings from study one indicate that measurement
incomparability may be greater among groups that are not defined by manifest variables such as gender, language and ethnicity. Findings from study two highlight the importance of examining within group heterogeneity in measurement comparability analyses; otherwise, violations of the WGH assumption may lead to grossly overestimating comparability claims.
PREFACE

UBC Research Ethics Certificate of Approval issued by the Behavioural Research Ethics Board. Certificate Number: H10-01566
# TABLE OF CONTENTS

Abstract ........................................................................................................................................... ii

Preface ................................................................................................................................................ iv

Table of Contents .......................................................................................................................... v

List of Tables ...................................................................................................................................... ix

List of Figures ..................................................................................................................................... xi

Acknowledgments ........................................................................................................................... xii

Dedication .......................................................................................................................................... xiii

1. Introduction .................................................................................................................................... 1

1.1. Research Studies ......................................................................................................................... 4

1.1.1. Study One: “Investigation of Within Group Heterogeneity in DIF Analyses Using International Large-Scale Assessment Data” ......................................................................................... 4

1.1.2. Study Two: “Effects of Within Group Heterogeneity on Methods Used to Detect Manifest DIF” ........................................................................................................................................... 5

1.1.3. Connection among the Two Studies ....................................................................................... 6

1.2. Manifest DIF Detection Methods ............................................................................................... 7

1.2.1. The Mantel-Haenszel Method of DIF Detection ................................................................... 8

1.2.2. The Unidimensional Item Response Theory Method of DIF Detection .............................. 10

1.2.3. The Logistic Regression Method of DIF Detection ............................................................... 12

1.2.4. The Assumption of Within Group Homogeneity .................................................................. 13

1.3. Latent Variable Mixture Models ............................................................................................... 15
1.3.1. The Discrete Mixture Distribution Model............................................................... 17

1.4. Summary ..................................................................................................................... 20

2. Investigation of Within Group Heterogeneity in DIF Analyses Using International Large-Scale Assessment Data ......................................................................................... 21

2.1 Introduction .................................................................................................................. 21

2.2. The Present Study ..................................................................................................... 25

2.3. Method ......................................................................................................................... 26

2.3.1. Measure................................................................................................................... 26

2.3.2. Assessment Framework ......................................................................................... 29

2.3.3. Data ......................................................................................................................... 32

2.3.4. Background Questionnaires .................................................................................. 36

2.3.5. Comparison Group and Samples ............................................................................ 38

2.4. Analyses ....................................................................................................................... 40

2.4.1. Analysis of Unidimensionality .............................................................................. 41

2.4.2. Analysis of DIF ....................................................................................................... 42

2.4.3. Detection of DIF Using a Manifest Approach ....................................................... 43

2.4.4. Identification of DIF Using a Latent Class Approach .......................................... 44

2.4.5. Investigation of Sources of DIF ............................................................................ 45

2.4.6. Investigation of the Composition of the Resulting Latent Classes ....................... 46

2.5. Results ......................................................................................................................... 46

2.5.1. Results of Unidimensionality Analyses ................................................................. 46

2.5.2. Results of Manifest DIF Analyses ........................................................................ 47

2.5.3. Results of Latent Class DIF Analyses ................................................................. 49
2.5.4. Results of Substantive Analyses................................................................. 53
2.5.5. Results of Analyses of Latent Class Composition................................. 57
2.6. Discussion ....................................................................................................... 62

3. Effects of Within Group Heterogeneity on Methods Used to Detect Manifest DIF ........ 68
3.1. Introduction..................................................................................................... 68
3.2. Literature Review ........................................................................................ 69
3.2.1. Investigation of the WGH Assumption in Real Data .............................. 69
3.2.2. Simulation Studies Examining Within Group Heterogeneity ................. 70
3.3. The Current Study ....................................................................................... 72
3.4. Method .......................................................................................................... 73
3.4.1. Data Generation ..................................................................................... 73
3.4.2. Simulation Factors .................................................................................. 77
3.4.3. Simulation Procedures .......................................................................... 79
3.4.4. Analysis of Simulated Data ................................................................. 83
3.4.5. Analysis of Manifest DIF ...................................................................... 83
3.5. Results .......................................................................................................... 84
3.5.1. The 0% DIF Condition ........................................................................ 84
3.5.2. False Positive DIF Detection Rates .................................................... 84
3.5.3. Correct DIF Detection Rates ............................................................... 87
3.6. Summary and Discussion ......................................................................... 91

4. Conclusion ....................................................................................................... 94
4.1. Contributions ............................................................................................... 94
4.2. Implications ................................................................................................. 96
4.3. Limitations .................................................................................................. 97
4.4. Future Research...................................................................................................................................... 98

References.................................................................................................................................................. 100

APPENDIX A: PIRLS 2006 Reader Passages and Items.............................................................................. 114

APPENDIX B: Examining Class Composition Using a Multinomial Multilevel Logistic Regression Model .................................................................................................................................................. 135
LIST OF TABLES

Table 1.1 Data for the \( k \)th Matched Set of Reference and Focal Group Examinees ......................... 9
Table 2.1 Distribution of Literary and Informational Blocks across Booklets ................................. 32
Table 2.2 Reader Booklet Items (Literary Experience) ..................................................................... 34
Table 2.3 Reader Booklet Items (Reading for Information) ............................................................. 35
Table 2.4 Reader Booklet Descriptive Statistics .................................................................................. 40
Table 2.5 Goodness-of-Fit Criteria for Evaluating Unidimensionality .................................................. 47
Table 2.6 Pairwise Country Comparisons ............................................................................................. 48
Table 2.7 Fit Statistics for the PIRLS 2006 Reader .............................................................................. 50
Table 2.8 Effect Size of DIF Identified Using a Latent Class DIF Approach ........................................... 51
Table 2.9 Correspondence between Items Identified by Latent Class and Manifest DIF Approaches .................................................................................................................................................. 52
Table 2.10 Grouping of Manifest DIF Items .......................................................................................... 55
Table 2.11 Grouping of Latent Class DIF Items ..................................................................................... 57
Table 2.12 Average Plausible Values Associated with Examinees Country of Residence and LC 58
Table 2.13 Results of Multinomial Logistic Regression Analyses ....................................................... 60
Table 2.14 Discriminant Function Analysis Structure Matrix ............................................................. 62
Table 3.1 Overview of Manipulated IVs ................................................................................................. 79
Table 3.2 Item Parameters ...................................................................................................................... 82
Table 3.3 Results of Mean Percentage False Positive Rates Using the MH Method of DIF Detection .................................................................................................................................................. 85
Table 3.4 Results of Mean Percentage False Positive Rates Using the LR Method of DIF Detection .................................................................................................................................................. 86
Table 3.5 Results of Mean Percentage Correct Detection Rates Using the MH Method of DIF Detection .................................................................................................................................................. 88

Table 3.6 Results of Mean Percentage Correct Detection Rates Using the LR Method of DIF Detection .................................................................................................................................................. 89

Table B.1 Number of Level-1 & Level-2 Predictors and Average Number of Examinees per Classroom .................................................................................................................................................. 136

Table B.2 Results of Multinomial Multilevel Analyses .................................................................................................................................................. 138
LIST OF FIGURES

Figure 1.1 ICC for Males versus Females ................................................................. 12

Figure 2.1 Discriminant Functions at Group Centroids ........................................... 61

Figure 3.1 0% Within Group Heterogeneity ............................................................ 74

Figure 3.2 Simulation of 20% to 80% Within Group Heterogeneity Conditions ........... 76
ACKNOWLEDGMENTS

I would like to start by expressing my deepest gratitude to my supervisor Dr. Kadriye Ercikan for her committed support and guidance throughout my doctoral studies. It is undoubtedly with her care and dedication that I was able to overcome the challenges of writing and defending my dissertation and to publish well beyond it. I would also like to thank Dr. Bruno Zumbo for his support, guidance, and dedication to my work as student and scholar. It is through the rich and enlightening discussions with Drs. Ercikan and Zumbo that I complete my doctoral studies with a myriad of ideas for future endeavours to pursue and research. I would also like to thank Dr. Matthias von Davier for his support, guidance and thorough discussions on mixture models. His advice and feedback greatly facilitated my understanding of mixture models and DIF in composite populations. I am also thankful to Dr. Jennifer Shapka for her initial input in my dissertation and to Dr. Nand Kishor for his suggestion to transfer to the Measurement, Evaluation and Research Methodology program.

I would like to thank my friends: Dallie, Rubab, Brent and Danielle for their continuous support, encouragement and friendship through the years. I would also like to thank my parents for instilling in me a passion for learning and growing. In closing, I would like to thank Gianfranco for his dedication, support, understanding and love.

I would also like to acknowledge the Social Sciences and Humanities Research Council for administering the doctoral fellowship that supported this research.
DEDICATION

To Giulia and Gianfranco, my two sources of inspiration.

To my parents for instilling in me a love of learning from a very young age
1. Introduction

In today’s highly competitive and globalised economy one of society’s key assets is a high quality education system. Data collected from educational large-scale assessments (ELSAs) are central to informing educational policies, and instructional practices. These assessments are also used to make inferences regarding the distribution of examinee competencies in policy relevant groups. Essential to making valid inferences from these assessments is to ascertain whether the assessments measure students’ ability in comparable ways and whether inferences made from ELSAs apply to the student groups targeted by the assessments. More specifically, the validity of inferences from ELSAs needs to be ascertained for all groups taking the test, including students from diverse linguistic, cultural and socioeconomic backgrounds, before test results can be used for decision-making (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999, p. 74).

Analyses of group comparability examine whether the assessments used to measure examinees’ knowledge and competencies assess the same constructs across groups being compared, and provide measures on the same scale that have similar levels of uncertainty. The importance of examining comparability is highlighted in the Standards for Educational and Psychological Testing, which states that empirical investigations should be conducted to examine whether examinees of equal knowledge and competency levels, irrespective of group membership (e.g., based on gender, ethnic, or linguistic backgrounds), are assigned (on average) the same test score (AERA, APA, NCME, & 1999, p. 74).

The analysis of comparability is typically conducted by investigating whether particular items or the entire test function differentially across compared groups of equal ability. At the test-level, comparability analyses examine whether score distributions differ across examinee
groups matched on ability whereas analyses at the item-level investigate whether compared examinee groups have the same probability of correctly responding to an item or set of items on a given test. When items are found to function differentially across examinee groups, or if differential item functioning (DIF) is found, panels of experts are typically used to further investigate whether or not an item displaying DIF is biased based on item content, language, and presentation.

In multi-lingual testing contexts, such as those in Canada, wherein tests are administered in English and French and in international assessments, wherein assessments are administered in over 40 languages, there are multiple factors that may affect construct and score comparability. These factors include: test-takers’ competency levels in the language of the test, familiarity with test content and format as well as linguistic, cultural and curricular differences among others. These factors can influence comparability within (and across) countries and across subpopulations such as the ones defined by gender or language background. Additional factors, such as the adaptation of measures to different languages and the inclusion of idiomatic expressions, vocabulary and words that are of differential difficulty for the various populations taking the assessment may also lead to DIF (Allalouf, 2003; Ercikan 2002; Ercikan & Koh, 2005; Oliveri & Ercikan, 2011).

Several studies have been conducted to investigate the presence and sources of DIF as well as the accuracy of various methods used to examine it. Primarily, these studies have been conducted using manifest DIF detection methods, such as Mantel Haenszel (MH) (Holland & Thayer, 1988), logistic regression (LR) (Swaminathan & Rogers, 1990), and unidimensional item response theory (IRT) (Hambleton, Swaminathan, & Rogers, 1991; Muraki, 1992). These three methods will be reviewed in later subsections of this chapter. The reader is directed to
Camilli and Shepard (1994), Clauser and Mazor (1998), and Holland and Wainer (1993) for additional reviews.

Manifest DIF detection methods have the underlying assumption of within group homogeneity (WGH). The WGH assumption denotes that claims based upon comparing examinee groups (e.g., boys vs. girls) generalize across all examinees within the group (e.g., all girls). This type of claim implicitly suggests that there is no statistically significant within group variability and that “all girls” or “all boys” within the examined sample are similar with respect to their response patterns.

Findings from previous studies (Cohen & Bolt, 2005; De Ayala, Seock, Stapleton, & Dayton, 2002) suggest that the assumption of WGH may not hold because varying proportions of examinees within a manifest group (e.g., subgroups of girls) may respond to a test in ways that diverge from the overall gender group (e.g., all girls). Consequently, the assumption of population homogeneity may lead to results that are neither generalizable nor applicable to test-taker subgroups (i.e., those that differ from their own manifest group). As generalizability of research findings is one of the key criterion researchers use to determine the value and usefulness of research, this is an important issue that needs to be addressed (Ercikan, 2009, p.211).

This research investigates the broad concern of the violation of the WGH assumption in manifest group DIF detection methods. To this end, two studies were conducted. The next two chapters describe each of the two studies conducted to investigate the WGH assumption associated with manifest DIF detection methods. The last chapter is a discussion of conclusions wherein findings from the two studies, implications and contributions of this research are discussed.
This introductory chapter has four sections. The first section summarizes the two studies that were conducted to investigate the WGH assumption associated with manifest DIF methods and discusses four ways in which these studies were interconnected. The second section describes manifest DIF methods, related mathematical models and the underlying assumption of WGH. Next, a description of latent variable mixture models which empirically evaluate the WGH assumption is given. The last section briefly summarizes this chapter.

1.1. Research Studies

1.1.1. Study One: “Investigation of Within Group Heterogeneity in DIF Analyses Using International Large-Scale Assessment Data”

The first study is entitled: “Investigation of Within Group Heterogeneity in DIF Analyses Using International Large-Scale Assessment Data.” It investigated the WGH assumption associated with manifest DIF methods using real data from an international measure of reading literacy; namely, the Progress in International Reading Literacy Study, 2006 (PIRLS, 2006) administered to examinees from four non-Indo European language speaking countries: Kuwait, Qatar, Hong Kong and Chinese Taipei.

International assessments, such as PIRLS are administered across multiple countries and languages to collect information regarding how the academic competencies of students from one country compare to that of students from other countries (Cook, 2006). Data from an international ELSA was used because as compared to tests administered to examinees that speak the same language or come from the same country, examinees participating in international ELSAs may be more diverse as they come from different countries and cultures, speak different languages and have received diverse instruction and exposure to the curriculum (Oliveri & von Davier, 2011). The tenability of the WGH assumption within this context is thus contested.
In this study, DIF was investigated using two approaches: a manifest and a latent class DIF approach. Manifest DIF was investigated using LR (Swaminathan & Rogers, 1990) and unidimensional IRT (Hambleton et al., 1991; Muraki, 1992) manifest DIF detection methods. Latent class DIF was investigated using the discrete mixture distribution (DMD) model (von Davier & Yamamoto, 2004). This study examined the following three research questions:

1) What is the degree of group homogeneity for the four selected countries with respect to the underlying structure of the response patterns?

2) To what extent do manifest and latent class DIF detection methods result in the identification of the same DIF items?

3) What are the sources of incomparability underlying differentially functioning items?

The first question examined the homogeneity of response patterns within groups and the number of latent groups or classes in the underlying structure of the data. The last two questions examined key aspects of measurement comparability research: investigation of DIF and its sources using two different (manifest and latent class) DIF approaches. Sources of DIF need to be investigated in order for DIF detection to be used to minimize bias or provide information about student response patterns.

1.1.2. Study Two: “Effects of Within Group Heterogeneity on Methods Used to Detect Manifest DIF”

The second study is entitled “Effects of Within Group Heterogeneity on Methods Used to Detect Manifest DIF.” This study used simulated data to examine the effects of within group heterogeneity on accuracy of DIF detection. Accuracy was investigated by examining false positive (FP) and correct detection (CD) rates associated with manifest DIF methods; namely, the LR and MH DIF methods. Simulated data was used because such a design allowed for
controlling particular factors such as amount of DIF as well as the degree of within group heterogeneity in the data.

In the simulation study, sample sizes and item characteristics were designed to illustrate the context of international ELSAs wherein compared examinee groups may come from diverse linguistic or cultural backgrounds in different proportions. Moreover, item parameters from study 1 were used for a proportion of simulated test items.

1.1.3. Connection among the Two Studies

The two studies conducted for this dissertation were connected in three ways. First, a common issue was investigated across the two studies: the effects of violating the WGH assumption on the accuracy of manifest DIF detection. Second, bridging the two studies is the use of a common international large-scale assessment context: the first study used data from an international ELSA and the second was designed to exemplify a context occurring frequently in international ELSAs (i.e., comparing culturally or linguistically diverse examinee groups). Third, both studies investigated violation of WGH on DIF as identified by the LR DIF method. The two studies jointly addressed the paucity in measurement comparability research associated with examining the effects of within group heterogeneity when using manifest DIF detection methods. This gap in the literature was highlighted by Rutkowski, Gonzalez, Joncas and von Davier (2010):

“The technical complexities and sheer size of international large-scale assessment databases often cause hesitation on the part of the applied researcher interested in analyzing them. Further, inappropriate choice or application of statistical methods is a common problem in applied research using these databases.” (p.142)

Hence, due to the size and complexities of international large-scale assessment data, applied researchers often detract from using these data leading to a paucity of research in this area.
1.2. Manifest DIF Detection Methods

Manifest DIF detection methods were developed during a high-stakes testing context wherein tests were used to determine whether individual examinees would graduate, be admitted to universities or receive scholarships. Within this context, these methods were developed for the purposes of detecting items that may have unfairly (dis)advantaged a group of examinees (e.g., Hispanics) as compared to another group (e.g., examinees who are White) that were matched on ability (Holland & Thayer, 1988; Zumbo, 2007a). In DIF parlance, the latter group (typically favoured by the item) is referred to as the reference group and the former (the group often disadvantaged by the item) is described as the focal group.

Manifest DIF detection methods can be categorized into three main groups: (1) methods based on the use of contingency tables (e.g., MH; Holland & Thayer, 1988), or regression models (e.g., LR; Swaminathan & Rogers, 1990), (2) unidimensional IRT (Hambleton, Swaminathan, & Rogers, 1991; Muraki, 1992) models, and (3) models that view DIF as occurring due to multidimensionality or secondary nuisance dimensions and factors that are not associated with the construct being measured, such as Simultaneous Item Bias Test (SIBTEST) (Roussos & Stout, 1996; Shealy & Stout, 1993).

Detection of DIF using a manifest approach typically requires three steps. In the first step, examinees are grouped using a manifest variable selected a priori (e.g., gender). Second, examinees are matched on ability using a total latent or observed score. The LR and MH methods use the total score summed over items to match examinees whereas IRT and SIBTEST use a total latent score; that is a score that is not directly observable but rather is computed from student responses to test items (Clauser & Mazor, 1998). The third step involves comparing the likelihood of examinees matched on ability to score correctly on an item; such comparisons are conducted in different ways depending on the method used.
Some of the most commonly used methods employed to detect manifest DIF include MH, LR and unidimensional IRT. These methods were used in this dissertation. In the first study, LR and unidimensional IRT were used because they allow for estimation of DIF in item types (dichotomous and polytomous items) and types of DIF (uniform and non-uniform DIF) that are likely to occur in multi-language and international ELSAs (Oliveri & Ercikan, 2011). Uniform DIF occurs when one group is favoured over the other consistently throughout the various points of the scale. Non-uniform DIF occurs when one group is disadvantaged at some parts of the scale and advantaged at other points (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990). As the effects of language on performance may differ along the ability scale in multilingual assessments, the use of DIF detection methods that identify uniform and non-uniform DIF is important for examining DIF in language groups (Oliveri & Ercikan, 2011). In the second study, the MH and LR methods were used as examples of the most widely used manifest DIF detection methods. Also, these methods were analyzed using SPSS, which lent itself to a simulation study due to its syntax writing module. A latent variable mixture model was used in study one to investigate latent class DIF.

1.2.1. The Mantel-Haenszel Method of DIF Detection

In the MH method (Holland & Thayer, 1988), DIF is identified when different examinee groups who are matched on observed total score have differing likelihoods of scoring correctly on an item (Clauser & Mazor, 1998). $K$ two-by-two contingency tables (where $K$ is the number of score categories) are constructed for each test score to compare the probability of correct response associated with the number of examinees from the reference (R) and focal (F) groups. In the contingency table illustrated in Table 1.1, the total number of examinees in the former (R) group is indicated by $n_{Rk}$ and the total number of examinees from the latter (F) group by $n_{Fk}$. The total number of examinees who answered the items correctly is indicated by $m_{1k}$ and the total number
of examinees who answered the items incorrectly is represented by $m_{0k}$. $A_k$ and $C_k$ represents the number of examinees who answered the items correctly in the R and F groups, respectively, whereas $B_k$ and $D_k$ are the number of examinees who answered the item incorrectly in the R and F groups, respectively. $T_k$ is the total number of examinees.

**Table 1.1**

*Data for the $k^{th}$ Matched Set of Reference and Focal Group Examinees*

<table>
<thead>
<tr>
<th>Group</th>
<th>Score on Studied Item</th>
<th>$A_k$</th>
<th>$B_k$</th>
<th>$C_k$</th>
<th>$D_k$</th>
<th>$n_{Rk}$</th>
<th>$n_{Fk}$</th>
<th>$T_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$m_{1k}$</td>
<td>$m_{0k}$</td>
<td></td>
</tr>
</tbody>
</table>

Analysis of DIF using the MH method results in a $\chi^2$ distributed statistic with one degree of freedom. The $\chi^2$ statistic is obtained as shown in the equation below (Zwick & Ercikan, 1989, p.58):

$$MH\chi^2 = \frac{(\sum_k A_k - \sum E(A_k)) - \frac{1}{2})^2}{\sum_k \text{var}(A_k)}$$

Where the expectation for $A_k$ is calculated using the marginal frequencies, as below:

$$E(A_k) = n_{Rk}m_{1k}/T_k.$$
And the variance is found using the following equation:

$$Var(A_k) = \frac{n_Rk n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}.$$ 

To indicate magnitude of DIF, Mantel and Haenszel provided the estimator (\(\hat{\omega}\)). This estimator represents a common odds ratio or a difference measure and is calculated using the equation below:

$$\hat{\omega} = \frac{\sum A_k D_k / T_k}{\sum B_k C_k / T_k}.$$ 

Typically, the natural log of (\(\omega\)) is taken so the index is on a symmetric scale with a mean of zero. When this is done, positive values indicate that the reference group is being favoured by the item (Zwick & Ercikan, 1989).

1.2.2. The Unidimensional Item Response Theory Method of DIF Detection

In unidimensional IRT, the probability of scoring correctly on an item is modeled as a function of a person or ability parameter characterized by an S-shaped item characteristic curve (ICC). The ICC provides information regarding various aspects of an item, including the degree of difficulty associated with correctly endorsing an item, the degree to which the item discriminates among persons along different areas of the ability continuum and the degree of guessing associated with correctly responding to an item on the test (Hambleton et al., 1991; Lord, 1980; Zumbo, 1999).

Mathematically, the probability that a person with a score of \(\theta\) responds correctly to item \(i\) in the 3 parameter logistic (3PL) model (Lord, 1980) is shown below (Hambleton et al., 1991, p.17):
In this model, $a_i$ is the item discrimination parameter representing the degree to which the item discriminates between persons in the latent ability continuum. $b_i$ is the item difficulty parameter indicating the degree of the latent variable (ability) needed to score correctly on the item. $c_i$ is the pseudo-chance-level parameter capturing the probability of examinees with very low ability responding to an item correctly. The 2PL model is equivalent to the 3PL model when the guessing parameter has been set to 0 (Hambleton et al., 1991). The 1PL (Rasch) model is used for models that use only item difficulty parameters to depict the probability of examinees’ correctly answering an item and examinees’ ability.

There are various approaches to identifying DIF using a unidimensional IRT model. One way is to compare item parameters across focal and reference groups matched on latent ability score to determine whether differences in item parameters are statistically significant. Another way is to compare the ICCs of the focal versus the reference group. If the ICCs are identical for the group, the item does not display DIF. On the other hand, if statistically significant differences between ICCs are present, the item is identified as DIF. To illustrate, Figure 1.1 shows two monotonically increasing ICCs depicting uniform DIF across female versus male students, displayed by the dashed and solid curves, respectively. Uniform DIF favouring females is indicated because females are consistently favoured along all points of the ability scale (x-axis) as compared to males who are disfavoured throughout the ability scale.
Figure 1.1. ICC for Males versus Females

1.2.3. The Logistic Regression Method of DIF Detection

The LR DIF detection method is another commonly used method to detect DIF. As in IRT, the LR method can be used to graphically display the relationship between examinees’ total test score and their probability of correctly responding to an item. Similarly, DIF is identified when differences are found between the probability of responding correctly for matched comparison groups. However, unlike IRT, which is based on the use of a latent score for matching examinees, the LR method uses a total observed score for matching examinees.

The LR method is based on a logistic regression equation, and involves the regression of the item score on the total test score, group membership (e.g., language, gender, ethnicity, or race), and the interaction between total test score and group membership. The regression equation is conducted in a step-wise fashion as the grouping variable and the interaction terms are added to the model resulting in a $\chi^2$ based test of significance and an effect size ($R^2$) statistic obtained for each step. The
LR method is used to identify DIF by comparing the response of different student groups (e.g., male vs. female student responses) and testing for significance of the $\chi^2$ statistic. Differences between $R^2$ values with the addition of each new step in the model provide a measure of variance accounted for by the added variable (Swaminathan & Rogers, 1990; Zumbo, 1999).

In the LR model (Swaminathan & Rogers, 1990, p.362), the probability of a correct response to an item is:

$$P (u_{ij} = 1 | \theta_{ij}) = \frac{\exp (\beta_{0j} + \beta_{1j} \theta_{ij})}{1 + \exp (\beta_{0j} + \beta_{1j} \theta_{ij})}, i = 1, \ldots, n, j = 1, 2.$$  

Where, $P (u_{ij} = 1)$ is the response of person $i$ in group $j$ to the studied item;

$\beta_{0j}$ is the intercept parameter;

$\beta_{1j}$ is the slope parameter for group $j$, and

$\theta_{ij}$ is the observed ability of person $i$ in group $j$.

In this model, items are identified as DIF-free when $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$ as the LR functions are considered to be the same for the comparison groups (Swaminathan & Rogers, 1990).

1.2.4. The Assumption of Within Group Homogeneity

The assumption of WGH in psychometric modeling was described over two decades ago in a presidential address given by Muthén at the 1989 Meeting of the Psychometric Society:

In interacting with substantive researchers during recent years on issues of psychometric modeling, I have encountered an interesting common theme,
namely that of population heterogeneity. Data are frequently analyzed as if they were obtained from a single population, although it is often unlikely that all individuals in our sample have the same set of parameter values (p. 558).

Currently, more than two decades later, the majority of measurement comparability studies using manifest DIF detection methods continue to be conducted using data as if it came from one single population.

Presently, a few studies (Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005; Sawatzky, Ratner, Johnson, Kopec, & Zumbo, 2009; von Davier & Yamamoto, 2004) have been conducted to empirically investigate this assumption. Findings from these studies suggest that this assumption may not be accurate when analyzing item response data from diverse populations because there is a great degree of heterogeneity within groups. Sources of group heterogeneity may include receiving different test preparation training, using diverse cognitive strategies when responding to tests, having different motivations for learning and having been taught using different instructional methods (von Davier & Yamamoto, 2004).

The use of manifest DIF detection methods when the within group response data are heterogeneous may have three consequences. One consequence is the reduction of power to correctly detect DIF. Specifically, in previous simulation studies conducted by De Ayala et al. (2002) and Samuelsen (2005) reduced correct DIF detection rates (to below nominal levels) were found when compared groups were designed to be heterogeneous. That is, two subgroups within each compared group were created by introducing DIF to only one of the two subgroups. In contrast, findings from previous simulation studies (Rudner, Getson, & Knight, 1980; Swaminathan & Rogers, 1990) showed that manifest DIF detection methods yield high correct DIF detection rates (above 90%) when homogeneous simulee groups were compared. A second consequence may be DIF cancellation, wherein the effect of DIF items favouring some examinees may be cancelled by having the same item
disfavour other examinees within the group (Cohen & Bolt, 2005). Third, if the sample is assumed to be homogeneous and analyses to examine its heterogeneity are not conducted, we may run the risk of obtaining findings that do not generalize to all examinees within the group. This may threaten the validity of inferences made based upon the assessment and lead us to findings that are of limited applicability, meaning and use (Zumbo, 2007b).

In brief, heterogeneity in students’ responses may be expected when administering international measures (e.g., the Programme for International Student Assessment or PIRLS) because these tests are administered to examinees from diverse cultural and ethnic backgrounds (Oliveri & von Davier, 2011). Due to this potential heterogeneity, alternative approaches to analyzing student response data should be sought. One approach to capturing such heterogeneity in examinees’ responding is the use of latent variable mixture models (Sawatzky et al., 2009, von Davier & Yamamoto, 2004, Zumbo, 2009).

1.3. Latent Variable Mixture Models

The use of latent variable mixture models allows for empirically examining the presence of LCs (groups of examinees with homogenous response patterns) when conducting analyses of DIF. Unlike the manifest DIF methods, the DIF analyses using latent class models do not assume homogeneity of item response patterns within set comparison groups (Sawatzky et al., 2009). The identification of DIF, in latent variable mixture models consists of three steps. First, examinees are assigned to LCs based on their response patterns to test items. Next, variables that may explain LC membership are investigated. Membership of examinees into classes is deemed latent because it is not pre-specified prior to conducting data analysis (as when manifest DIF detection methods are used) but rather it is inferred from the data based upon examinees’ responses using statistical analyses (von Davier, 2001). Third, the presence of DIF is investigated by comparing item difficulty parameters associated with each of the resulting LCs.
To elaborate, if we compared the performance of Asian versus African American students on an achievement test using a latent class DIF approach, we may obtain two LCs. One LC may contain 65% of Asians and 77% of African Americans, and the other LC may contain the remaining proportion of examinees. Each LC may have differing parameters depending on their responses to test items. Moreover, each LC may contain varying proportions of examinees from each ethnic background due to their similarity in item responding, which may be more highly influenced by factors such as type of school they attended, neighbourhood in which they lived or instruction they received than by examinees’ ethnic background.

As discussed in Rupp (2005), the use of latent variable mixture models allows for closer examination of alternative variables, such as socio-economic status background, perceived valence of task, educational level of parents or resources available at home which may serve to explain a greater proportion of variance in examinees’ item responses than other variables (e.g., gender or ethnic background). As discussed by Kelderman and Macready (1990), this may be because the use of latent grouping variables allows for the assessment of DIF without tying that DIF to any specific variable or set of variables before data analyses are conducted.

The most well-known early latent variable approaches used to examine item response data include the use of the latent class analysis (Haertel, 1989) and rule space methodology (Tatsuoka, 1983) models. A central idea unifying these approaches is that various items tap different skills or examinee attributes which can be examined empirically by generating a matrix of relations between the items in the test and the skills needed to respond to these items (von Davier, 2005a). Examinees are assigned to one of a number of (latent) classes depending on their item responses. Examinees within each of these various classes are viewed as having different probabilities of responding in particular ways, depending on the values of the skill variables defining the LC (Mislevy, Wilson, Ercikan, & Chudowsky, 2001).
Latent variable mixture models have recently been used in various fields including psychology (Maij-de Meij, Kelderman, & van der Flier, 2008) and education (Bolt, Cohen, & Wollack, 2001; Cohen & Bolt, 2005; Samuelsen, 2005; Webb, Cohen, & Schwanenflugel, 2008). Recent applications of latent variable mixture models in educational research include the analysis of data from ELSAs in reading (Webb et al., 2008), mathematics (Cohen & Bolt, 2005), and English language arts (Bolt et al., 2001). Findings from these studies suggest that the use of latent variable mixture models allow for more in-depth investigations of students’ strengths and weaknesses, strategies examinee subgroups use when solving test questions as well as particular content test-taker subgroups have learned or have yet to learn.

1.3.1. The Discrete Mixture Distribution Model

In this study, the Discrete Mixture Distribution (DMD) model (von Davier & Yamamoto, 2004) was used to investigate latent class DIF. The DMD model is a type of latent variable mixture model which extends the mixed Rasch model (von Davier & Rost, 1995) to multi-parameter mixture IRT (MIRT) models to model unobserved grouping variables. It also extends the mixed Birnbaum model which uses dichotomous variables (Smit, Kelderman, & van der Flier, 2000) to models for polytomous data. Thus, the DMD model allows for analyses of mixed item types (dichotomous and polytomous) typically found in ELSAs. In the DMD model (von Davier, 2008, p. 261), the probability of a correct response \( (x) \) is:

\[
P\left(X_i = x \left| \bar{\beta}_i, \bar{\alpha}_i, \bar{\gamma}_i, g \right. \right) = \frac{\exp[\beta_{xi}g + \Sigma_{k=1}^{K} \gamma_{ik}q_{ik}(a_k)]}{1 + \Sigma_{y=1}^{m_i} \exp[\beta_{yi}g + \Sigma_{k=1}^{K} \gamma_{yik}q_{ik}(a_k)]}
\]
Where, subscript $x$ indicates a correct response, $i$ indicates an item, $g$ refers to the LC (or examinee group), and $k$ indicates the number of skill levels needed to respond correctly to the item. In the simplest (dichotomous) case there are two $k$ levels (mastery vs. non mastery) of the measured learning outcome.

$\beta$ are item difficulty parameters for the dichotomous case and threshold parameters for the polytomous case;

$\gamma$ are slope (or discrimination) parameters;

$q_{ik}$ are $Q$ matrix entries which relate item $i$ to skill $k$. If skill $k$ is needed for item $i$, then $q_{ik} > 0$ and 0 otherwise;

$\theta$ denotes a latent variable depicting skills needed to correctly respond to the item;

$a_k$ refers to the levels of the skills needed to respond correctly;

$\tilde{a}^*$ represents user specified skill levels needed to represent skill profiles;

And arrows indicate parameter vectors.

In the DMD model, the marginal probability of a response vector is:

$$P\left(\tilde{X} = \tilde{x} \left| \tilde{\beta}, Q, \tilde{\gamma} \right.\right) = \sum_g \pi_g \sum_{\tilde{\alpha}} \pi_{\tilde{\alpha}|g} \prod_{i=1}^I P \left( X_i = x \left| \tilde{\beta}_i, \tilde{\alpha}, q_{i}, \tilde{\gamma}_i, g \right. \right).$$

with cube-valued (classes $g$ * items $i$ * skills $k$) item difficulty parameters

$$\tilde{\beta} = (\beta_{gik})_{g=1..g,i=1..I;k=1..K},$$
and cube-valued (classes \( g \) * items \( i \) * skills \( k \)) slope (discrimination) parameters

\[
\vec{y} = (\gamma_{gik})_{g=1..g; i=1..I; k=1..K},
\]

Where \( \pi_g \) denotes the size of the latent class; and

\( \pi_{a|g} \) denotes the class specific distribution of skill patterns \( \vec{a} \) in group \( g \).

The application of the DMD model to national ELSAs was demonstrated in a study conducted by von Davier and Yamamoto (2004) which used data from a national large-scale assessment of mathematics administered to 15 year-old students to investigate whether all items in the measure function equally across all subpopulations. Findings from the study conducted by von Davier and Yamamoto (2004) revealed that improved fit was obtained when heterogeneity (in the form of three subpopulations) was allowed for the data structure; that is, the data was not constrained to be homogeneous as when manifest DIF detection methods are used. Moreover, findings also suggest that rather than constraining all item parameters to be equal across the three subpopulations examined in their study, to accommodate empirical evidence that some items function differently in the components of the composite population, the difficulty and other parameters of a set of items should be allowed to differ across groups of subpopulations. These findings suggest that the use of more flexible models (e.g., those that allow for subpopulations in the data) can yield more accurate results regarding DIF than models that assume group homogeneity.

Although the latent variable mixture model (e.g., the DMD model) is promising for investigating DIF, this approach has not yet been extensively applied in the investigation of DIF. As highlighted by Ferne and Rupp (2007) in language testing:
Latent class and mixture analyses allow the detection of unobserved heterogeneity in data sets for DIF analyses (i.e., they allow analysts to detect the groups of learners from the data themselves without explicit prior specification), but those analyses are rarely used in language testing at this point in time (p. 115).

As noted, an overarching intention of this dissertation was to conduct two studies that will contribute towards filling this gap in the literature.

1.4. Summary

This chapter introduced key issues associated with examining measurement comparability of ELSAs and the manifest DIF detection methods (i.e., MH, LR, and unidimensional IRT) traditionally used to investigate comparability at the item-level. This chapter also described the WGH assumption underlying the use of manifest DIF methods as well as the shortcomings associated with assuming WGH without empirically evaluating the test data structure. Two limitations were primarily highlighted; these included (1) limited generalizability of findings to subgroups of students who may differ from the group being analyzed, and (2) reduction in power to detect DIF. As an alternative, latent variable mixture models were proposed. These models allow for the identification of homogeneous subpopulations potentially present within a composite population. In particular, the DMD model was suggested because it is a flexible model which allows for the analysis of dichotomous and polytomous item types frequently found in ELSAs. These models were also suggested because they may yield improved model-data fit estimates with data that is heterogeneous, lead to results that are more applicable to subpopulations as well as allow for more thorough investigations of background variables underlying DIF.
2. Investigation of Within Group Heterogeneity in DIF Analyses Using International Large-Scale Assessment Data

2.1. Introduction

Educational large-scale assessments (ELSAs) are administered at various levels including provincial, national, and international levels. Two ELSAs are administered in North America. The National Assessment of Educational Progress (NAEP) administered in the United States and the Pan-Canadian Assessment Program (PCAP) administered in Canada. At the international level, ELSAs include the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS). At these various levels, results from ELSAs are used to make inferences regarding the distribution of examinee competencies in policy relevant groups and to inform policies, curriculum planning and educational decision-making.

International assessments, in particular, are administered across multiple countries and languages to collect information regarding how the academic competencies of students from one country compare to those of students from other countries (Cook, 2006). To illustrate, PIRLS 2006 was translated into 44 different languages, and was administered across 40 countries and 5 Canadian provinces (Martin, Mullis, & Kennedy, 2007), and PISA 2006 was translated into 42 languages and had 81 national versions (Turner & Adams, 2007).

Central to using results from international ELSAs to make comparisons across groups is to establish measurement comparability across test forms. Analyses of measurement comparability investigate whether assessments measuring examinees’ knowledge and competencies measure the same constructs across comparison groups, and yield measures on the same scale with similar levels of accuracy. These analyses are conducted to investigate whether examinee groups of equal ability obtain equivalent scores irrespective of gender, ethnic, or linguistic backgrounds. Such investigations
entail examining whether particular items or the entire test function differentially across comparison
groups.

Analyses conducted to investigate test-level comparability examine whether test scores differ
across comparison groups matched on ability (American Educational Research Association [AERA],
American Psychological Association [APA], National Council on Measurement in Education
[NCME], & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 1999).
Test-level analyses of comparability are typically conducted using confirmatory or exploratory factor
analyses (Oliveri & Ercikan, 2011) or comparisons of test characteristic curves (Ercikan & Gonzalez,
2008). In addition, differential test functioning can be analyzed using methods such as logistic
regression, parametric and non-parametric IRT. Please see Oliveri, Olson, Ercikan and Zumbo (in
press) for a more thorough discussion of differential test functioning and a description of methods
used to analyze it.

Item-level analyses are conducted to examine whether examinees of equal ability have the
same probability of correctly responding to an item or set of items on a given test. When items are
found to function differentially across examinee groups, or if differential item functioning (DIF) is
found, further analyses are typically conducted by using panels of experts to identify potential sources
of DIF and decide whether an item displaying DIF is biased. The presence of items that function
differentially in a test may affect the performance of diverse examinee groups, and threaten
researchers’ and test users’ ability to use test scores to make valid inferences about student
performance (AERA, APA, & NCME, 1999). More specifically, DIF threatens the comparability of
items on a measure and may lead to interpreting sources of DIF that are irrelevant to the construct
being measured, such as lack of clarity in test instructions or differences in vocabulary difficulty, as
indicative of true performance differences among groups (Sireci, 1997).
Previous research has demonstrated that results from multilingual versions of assessments cannot be assumed to be comparable (Allalouf, Hambleton, & Sireci, 1999; Cook, 2006; Ercikan & McCreith, 2002; Gierl & Khaliq, 2001; Hambleton, Merenda, & Spielberger, 2005), and that such assessments may contain large percentages of differentially functioning items (Allalouf et al., 1999; Ercikan & McCreith, 2002; Gierl & Khaliq, 2001; Hambleton et al., 2005). For example, Allalouf et al., (1999) used the Mantel-Haenszel (MH) method to examine DIF and found up to 65% of DIF items in their comparison of Hebrew and Russian translations of the Psychometric Entrance Test; Sireci and Berberoglu (2000) used item response theory (IRT) based DIF and found 42% of DIF items in the English and Turkish versions of a course evaluation form. Similarly, large numbers of DIF items, up to 50%, were found in studies that used the IRT based Linn-Harnisch (L-H; Linn & Harnisch, 1981) and Simultaneous Item Bias Test (SIBTEST) DIF detection methods to compare English and French versions of Canadian assessments across subjects including mathematics, science, and reading (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004). Moreover, previous studies have revealed several sources of DIF, including diversity of examinees’ cultural backgrounds, test-takers’ competency levels in the language of the test, familiarity with test content and format as well as the inclusion of idiomatic expressions, vocabulary and words that are of differential difficulty for the various populations taking the assessment (Allalouf, 2003; Ercikan 2002; Ercikan & Koh, 2005; Gierl & Khaliq, 2001).

To date, the majority of studies examining the comparability of ELSAs have been conducted using manifest DIF detection methods (e.g., SIBTEST; Shealy & Stout, 1993; Stout & Roussos, 1999, MH; Holland & Thayer, 1988 and unidimensional IRT; Hambleton et al., 1991; Muraki, 1992). These methods assume that there is homogeneity within manifest groups. As described by De Ayala et al. (2002), this assumption suggests that “individuals within a manifest group are more similar to one another than they are to members of the other manifest group” (p. 247). As such, items detected
to function differentially across members of one manifest group (e.g., DIF items favouring girls) are thought of as favouring *all* group members (e.g., *all* girls) in the population rather than a subset of examinees who have homogeneous item response patterns. The within group homogeneity (WGH) assumption has been contested by previous studies (De Ayala et al., 2002; Cohen & Bolt, 2005; Webb et al., 2008) which investigated DIF using an alternative (latent class) DIF approach. In latent class DIF, examinee groups (or latent classes; LCs) are formed by clustering individuals that have similar item response patterns. This approach differs from a manifest DIF approach wherein group membership is specified using a manifest variable (e.g., gender or language of the test). Previous studies using a latent class DIF approach reveal that large proportions of examinees (e.g., girls) have response patterns akin to the other (gender) group. To illustrate, Cohen and Bolt (2005) conducted a study to investigate gender DIF using the likelihood ratio manifest DIF detection method using data from a university mathematics placement test. Subsequently, these authors re-examined test items using an exploratory mixture IRT 3 parameter-logistic (3PL) model to investigate latent class DIF and obtained two LCs. Each of the resulting LCs had varying proportions of males and females (e.g., one LC had 50.8% of females and 37.6% of males).

Similar findings were obtained in another study conducted by De Ayala et al. (2002) wherein data from a college qualifications test was used to compare DIF resulting from a latent class and a manifest DIF approach. Specifically, the MH manifest DIF method was used to investigate DIF using examinees’ ethnic background (e.g., African American versus Whites) as a manifest grouping variable. Subsequently, these authors conducted further DIF analyses using a latent class DIF analysis model; two LCs resulted from these analyses. Each of the resulting LCs had varying proportions of African and European Americans (e.g., one LC had 67.6% of examinees that were African American and 32.4% of examinees that were European American). Findings from these two studies suggest that
manifest groups have heterogeneous response patterns and that larger differences in item response patterns exist between groups that are neither defined by gender nor ethnicity.

2.2. The Present Study

This study had two purposes. The first objective was to examine the WGH assumption underlying manifest DIF detection approaches. The second goal was to investigate DIF and its sources using two approaches: a manifest and a latent class DIF approach. To this end, this study investigated three research questions using data from an international ELSA of reading literacy (i.e., the PIRLS 2006 Reader) administered to examinees from distinctly different language groups and countries: Chinese Taipei, Hong Kong, Kuwait and Qatar. Within this context, the following three research questions were examined:

1) What is the degree of WGH for the four selected countries with respect to the underlying structure of the measure?

2) To what extent do manifest and latent class DIF detection methods result in the identification of the same DIF items?

3) What are the sources of incomparability underlying differentially functioning items?

The first question evaluated the degree to which the (manifest) comparison groups had homogeneous response patterns. These analyses were essential in determining the degree to which the WGH assumption associated with manifest DIF approaches is violated as well as determining the LCs within groups. The second and third questions examined two central aspects of measurement comparability research (i.e., investigation of DIF and its sources) using two different DIF analyses approaches: a manifest and a latent class DIF approach.
2.3. Method

2.3.1. Measure

This study used data from PIRLS 2006, an international measure of reading literacy led by the International Association for the Evaluation of Educational Achievement (IEA). This data was chosen for three reasons. First, the diversity of examinees participating in international assessments questions the tenability of the WGH assumption within this context. Second, the wealth of background questionnaire data available by using international measures allows for analyses of sources of DIF using various (e.g., teacher, home or school-related) variables. Third, the use of PIRLS 2006 data provides access to a large-scale data set necessary for statistical analyses.

International ELSAs (e.g., PIRLS) are administered to linguistically and ethnically diverse examinees from dozens of countries. These examinee groups are likely to differ in several ways including having diverse opportunities to learn, exposure to curricula and instructional approaches (Oliveri & von Davier, 2011; von Davier & Yamamoto, 2004). Despite the distinct populations participating in international ELSAs, the investigation of comparability within this context has for the most part been conducted using manifest DIF detection methods, which assume WGH. One objective of this study was thus to investigate the use of alternative (latent variable mixture) methods which do not assume sample homogeneity on the detection of DIF and its sources.

Moreover, in addition to examinee item response data, international ELSAs provide a wealth of data through their background questionnaire surveys, which allow for the investigation of factors underlying examinees’ response patterns. In this study, responses from the questionnaire administered to teachers were used to examine the influence of student and teacher background variables on examinees’ item response patterns.

Further, international ELSAs are administered to large groups of examinees (typically 4,000 examinees) per country. Large sample sizes are required for conducting DIF analyses using manifest
or latent class DIF approaches. For example, manifest DIF detection methods such as unidimensional IRT require sample sizes of at least 1,000 examinees when using a 3PL IRT model as was used in this study (Downing, 2003). Similarly, the use of latent variable mixture models also requires large sample sizes (e.g., a minimum within-class sample size of 200 examinees; Lubke & Neale, 2006).

PIRLS is conducted every five years to assess trends in and factors associated with reading literacy achievement: defined as the ability to understand and construct meaning from written language. Students read for various purposes including participation in the community, for enjoyment and for everyday purposes using diverse texts in print or in electronic forms such as books, magazines, newspapers, e-mail, and advertisements. PIRLS 2006 was administered to students who had four years of formal schooling with the intention to assess students who are older than 9.5 years; this age and grade were chosen because they mark the transition wherein students have typically learned to read, and are instead reading to learn.

PIRLS 2006 was administered to examinees from a total of 40 countries, 5 educational jurisdictions in Canada, and 2 educational systems in Belgium. The assessment was originally constructed in English and subsequently participating countries adapted the materials into their own target language(s) and cultural context. PIRLS 2006 was adapted into 44 languages. To ensure comparability, adaptations were verified through direct comparisons with the English (international) version (Martin et al., 2007). The assessments were developed jointly by the Reading Development Group, the PIRLS Reading Coordinator, the PIRLS 2006 National Research Coordinators (NRCs) from the participating countries, the PIRLS Item Development Task Force, and the staff from the TIMSS and PIRLS International Study Center. The development of PIRLS 2006 took approximately two years. It started in 2003 and lasted until August 2005 when the international versions of the assessment were ready for data collection. In early 2005, a field test was
conducted to investigate the measurement properties of items and passages across the participating countries. The NRCs were involved in developing test items and scoring-guides as well as approving reading passages to ensure selected passages were suitable for fourth-grade students with respect to content, interest and reading ability. NRCs also ensured that wherever possible, items did not make references to particular cultural groups.

PIRLS 2006 was composed of new passages and questions and also included passages and items from 2001 in order to measure trends across test administrations. Updates to PIRLS 2006 were intended to reflect findings from 2001 as well as developments in reading research since the development of the previous PIRLS assessment. Text organization and formats as well as passages and items included in PIRLS 2006 were selected from the types of materials likely to be read by Grade 4 students. Literary (e.g., short stories, narrative extracts, traditional tales, fables, myths, and play scripts) as well as informational texts (e.g., expository passages and persuasive writing) were included in PIRLS 2006 (Martin et al., 2007).

Student sampling for PIRLS 2006 consisted of a two-stage process. The first stage was a sampling of schools, and the second stage a sampling of intact classrooms from the target grade in the sampled schools. Most countries sampled approximately 4,000 students from about 150 schools. One or two intact classrooms from each school were sampled. Within the desired population, countries could define a population that excluded a small percentage (less than 5%) of certain kinds of schools or students that would be very difficult or resource intensive to test (e.g., schools for students with special needs or schools that were very small or located in remote rural areas). All countries examined in this study kept their excluded students below the 5 percent limit. The examined students had been taught by a single teacher in the assessment year.
2.3.2. Assessment Framework

PIRLS provides reports on how students in each country perform in reading comprehension overall and in particular measures the following three specific aspects of reading literacy: (1) purposes for reading, (2) processes for reading comprehension, and (3) reading behaviours and attitudes (Mullis, Kennedy, Martin, & Sainsbury, 2006).

The two purposes for reading are assessed in approximately equal proportions and are measured using separate scales: reading for literary experience and reading to acquire and use information. Reading for literary experience commonly refers to reading fiction or literary passages with items related to the plot, characters, events or settings of the story told in the voice of the narrator, or a principal character with events told chronologically or in varying time frames. Reading to acquire and use information refers to non-fiction reading with items related to facts or instructions, biographies or autobiographies where students are asked to read charts, graphs, compare and contrast information, or provide arguments for or against particular situations.

PIRLS assesses four increasingly more complex reading comprehension processes also measured in approximately equal proportions. These processes are: (1) retrieving ideas or information, (2) making inferences, (3) interpreting and integrating information and ideas, and (4) examining and evaluating text.

The process of retrieving information required students to make no or almost no interpretation of text as the meaning was evidently stated in the text. These questions required students to use various strategies to find, locate and comprehend explicitly stated content and to automatically understand and retrieve several pieces of information from text, including finding definitions for words or phrases, identifying the story setting or locating the topic sentence or main idea. Questions assessing students’ ability to make straightforward inferences required readers to connect two or more ideas, focus on global meaning or understanding of the whole text and make inferences about ideas.
not explicitly stated. Students were also asked to infer whether one event caused another, describe the relationship between two characters or identify generalizations made in the text. In PIRLS, these two processes were combined into the retrieval and making straightforward inferences scale which was designed to measure lower-order cognitive processes posing fewer cognitive demands on students and requiring minimum understanding or interpretation of text (Mullis, Martin, & Gonzalez, 2004).

Items measuring interpreting and integrating ideas and information asked students to comprehend text beyond the phrase or sentence level, focus on local or global meanings, and relate details to an overall theme and idea. Students were also asked to use their own perspective to answer questions and make connections on their prior or background knowledge of the world. In addition, students were asked to discern the overall message or theme of a text, consider alternatives to actions of characters, compare and contrast text information, infer the tone or mood of a story, or interpret real-world application of text information. Questions assessing students’ ability to examine and evaluate text required students to make comparisons across ideas or counter, question or confirm claims made in the text as well as to critically examine and evaluate meaning from the language of the text and related elements such as its genre and structure. Students were also asked to evaluate whether particular events were likely to happen, judge the completeness or clarity of information in the text, or identify the authors’ perspective. These questions required students to draw from their own perspectives and understanding of the world. These two processes were combined into the interpreting and integrating and evaluating scales. They measured higher-order cognitive complexity because answering related questions correctly required more in-depth interactions with the text (Mullis et al., 2004).

To cover the wide goals of the PIRLS assessment including providing information about key reading comprehension processes and reading purposes would have taken 7 hours of testing time per student; however, such demands on individual students would have been unreasonable. Therefore, to
reduce student testing time, matrix sampling was used. In matrix sampling, the total number of items is divided into a smaller number of items, which are administered to a subsample of examinees. Data are aggregated across subsamples and item subsets to obtain a measure of student group performance (AERA, APA, & NCME, 1999). In PIRLS 2006, the use of matrix sampling resulted in each student being tested for a total of 80 minutes with an additional 15-30 minutes to fill out the student questionnaire. The total number of passages and items were divided into ten 40-minute blocks of passages and items: 5 literary (labelled L1-5), and 5 informational (named I1-5) blocks, which were distributed across 13 booklets in a way that minimized the total number of pair combinations (see Table 2.1). Each student responded to one booklet which consisted of two 40-minute blocks of passages and items. As shown in Table 2.1, blocks L1-L4 and I1-I4 appeared three times across booklets. For example, L1 appeared with L2 in booklet 1, and with informational blocks I4 and I1 in booklets 8 and 9. The 13th booklet, the Reader consisted of blocks L5 and I5, which only appeared once; thus, to retain an equivalent proportion of randomly sampled examinees as other booklets administered in PIRLS, the Reader was administered to three times the number of examinees as any other booklet.
Table 2.1

Distribution of Literary and Informational Blocks across Booklets

<table>
<thead>
<tr>
<th>Booklet</th>
<th>Blocks</th>
<th>Booklet</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L1</td>
<td>8</td>
<td>I4</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td></td>
<td>L1</td>
</tr>
<tr>
<td>2</td>
<td>L2</td>
<td>9</td>
<td>L1</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td></td>
<td>I1</td>
</tr>
<tr>
<td>3</td>
<td>L3</td>
<td>10</td>
<td>I2</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td></td>
<td>L2</td>
</tr>
<tr>
<td>4</td>
<td>L4</td>
<td>11</td>
<td>L3</td>
</tr>
<tr>
<td></td>
<td>I1</td>
<td></td>
<td>I3</td>
</tr>
<tr>
<td>5</td>
<td>I1</td>
<td>12</td>
<td>I4</td>
</tr>
<tr>
<td></td>
<td>I2</td>
<td></td>
<td>L4</td>
</tr>
<tr>
<td>6</td>
<td>I2</td>
<td>13, Reader</td>
<td>L5</td>
</tr>
<tr>
<td></td>
<td>I3</td>
<td></td>
<td>I5</td>
</tr>
<tr>
<td>7</td>
<td>I3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generally, reading passages in the PIRLS 2006 measure consisted of approximately 1,000 words. PIRLS 2006 used two types of questions: multiple choice (MC), which were worth one point, and constructed-response (CR) questions, which were worth from one to three points, depending on the depth of understanding required to answer each CR question. In MC questions, students were asked to select one option from four response options, only one of which was correct. In CR questions, students were asked to construct an answer, often based upon their own knowledge and experiences. Each administered block of questions, provided on average 15 score points, consisting of approximately seven MC questions (1 point each), two or three short answer items (1 or 2 points each), and one extended response item (3 points).

2.3.3. Data

This study examined data from the 13th booklet (the PIRLS 2006 Reader) which was selected for two reasons. First, as it was administered to three times as many examinees as any other booklet, it provided large sample sizes required when conducting manifest and latent class DIF analyses.
Second, items from the Reader have been released to the public (see Appendix A; TIMSS, 2006). Access to all the items in the test is essential in identifying potential sources of DIF.

The Reader had a magazine-style format, and was intended to look authentic and parallel the reading choices Grade 4 readers would make. The Reader consists of two blocks of items. One 12-item block measuring literary experience called “Unbelievable Night”, and one 15-item block assessing informational knowledge named “Searching for Food.” These booklets are aligned with the content specifications outlined by PIRLS. In total, there are 27 items in the Reader, the 12-items in the literary block are evenly split between MC and CR items for a total of 16 points, and the 15-item informational block contains 7 MC and 8 CR items for a total of 18 points. Tables 2.2 and 2.3 show item numbers, item type and item scores, as well as the reading comprehension process assessed by each item for the literary experience and reading for information process booklets, respectively.
Table 2.2

*Reader Booklet Items (Literary Experience)*

<table>
<thead>
<tr>
<th>Reading Passage</th>
<th>Item #</th>
<th>Item Type</th>
<th>Reading Comp. Process</th>
<th>Maximum Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbelievable Night</td>
<td>1</td>
<td>MC</td>
<td>Focus and Retrieve</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>MC</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>MC</td>
<td>Interpreting</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>MC</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CR</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>CR</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>MC</td>
<td>Focus and Retrieve</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>CR</td>
<td>Interpreting</td>
<td>2 points</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>MC</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>CR</td>
<td>Focus and Retrieve</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>CR</td>
<td>Interpreting</td>
<td>3 points</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>CR</td>
<td>Examine and Evaluate</td>
<td>2 points</td>
</tr>
</tbody>
</table>

Total Score for Literary Block: 16 points

*Note:* Reading Comp. Process = Reading Comprehension Process

MC = Multiple Choice

CR = Constructed Response
### Table 2.3

**Reader Booklet Items (Reading for Information)**

<table>
<thead>
<tr>
<th>Reading Passage</th>
<th>Item #</th>
<th>Item Type</th>
<th>Reading Comp. Process</th>
<th>Maximum Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searching for Food</td>
<td>1</td>
<td>MC</td>
<td>Focus and Retrieve</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>MC</td>
<td>Focus and Retrieve</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>MC</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>MC</td>
<td>Interpreting</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CR</td>
<td>Interpreting</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>MC</td>
<td>Focus and Retrieve</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>CR</td>
<td>Examine and Evaluate</td>
<td>2 points</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>MC</td>
<td>Interpreting*</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>CR</td>
<td>Interpreting</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>CR</td>
<td>Interpreting</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>CR</td>
<td>Inference</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>CR</td>
<td>Interpreting</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>CR</td>
<td>Interpreting</td>
<td>2 points</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>MC</td>
<td>Examine and Evaluate</td>
<td>1 point</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>CR</td>
<td>Interpreting</td>
<td>2 points</td>
</tr>
<tr>
<td><strong>Total Score for</strong></td>
<td></td>
<td></td>
<td><strong>Information Block</strong></td>
<td>17 points**</td>
</tr>
<tr>
<td><strong>Total Combined</strong></td>
<td></td>
<td></td>
<td></td>
<td>33 points</td>
</tr>
</tbody>
</table>

*Note: Reading Comp. Process=Reading Comprehension Process

MC=Multiple Choice, CR=Constructed Response

* Question 8 was not administered to examinees from examined countries and was not included in any analyses

** Total excludes question 8
2.3.4. Background Questionnaires

To complement students’ reading achievement results, PIRLS administered questionnaires to students, parents, teachers, and school principals, which were designed to collect information about key factors related to students’ home and school environments. The student questionnaire collected information about resources students use and activities students engage in related to literacy both at school and at home. The learning to read survey asked parents or guardians to provide information related to literacy related resources they have at home as well as their perceptions of the support provided by the school and at home for literacy related activities. The school questionnaire gathered information from the school principal about the school’s demographics and the school’s reading curriculum and instructional policies. The curriculum questionnaire focused on the development and implementation of the reading curriculum of each of the participating countries or jurisdictions (Martin et al., 2007).

In conjunction with the item-response data from the PIRLS Reader, this study used data from the teacher questionnaire to investigate sources of DIF. This questionnaire was used because although students are exposed to a variety of literacy-related activities and experiences outside of school (e.g., at home and in the community), teachers play a critical role in the development of children’s reading. It is typically the teacher’s responsibility to structure the environment of the classroom as well as to select reading strategies and materials that s/he will use when teaching reading (Mullis, Martin, Kennedy, & Foy, 2007).

The teacher questionnaire was prepared by an expert panel that prepared specific questions that would be comparable across participating countries. NRCs provided particular formatting and wording changes that allowed for such comparability. Teacher questionnaires were field tested and changes included removing items, rewording or collapsing or expanding response categories based upon feedback from the field tests (Martin et al., 2007). These questionnaires were completed by each teacher instructing the sampled students; in cases when teachers taught more than one class, s/he
had to complete one questionnaire for each class. To link responses from the teacher questionnaire and the student cognitive data files, the student-teacher linkage data files provided by PIRLS were used (Foy & Kennedy, 2008).

The teacher questionnaire collected information about teacher demographics (e.g., teachers’ age, gender, and number of years teaching), class characteristics (e.g., number of years teachers taught a particular class, number of students in the class, as well as number of students with difficulty speaking the language of the test, needing enrichment or remediation) as well as degree of teacher training and their qualifications for teaching reading. The questionnaire also gathered data about teachers’ use of various instructional materials and their use of technology to teach reading (e.g., frequency of use of the school and/or classroom library and availability of computers in the class). Teachers also reported on instructional strategies and activities used in the classroom to teach reading (e.g., time spent on teaching language activities, type of activities used to teach students to decode, and comprehend passages read) as well as homework practices and home-school connections (e.g., how often teachers communicated with parents or guardians).

In total, the teacher questionnaire consisted of 42 questions. The following three questions were investigated in this study:

1) In a typical school week, what percentage of your time in class with students do you devote to working with individual students or small groups?

2) When you have reading instruction and/or do reading activities, how often do you create same-ability groups? (The following response options were available: Always or almost always, often, sometimes or never; teachers were asked to select only one option)

3) When you have reading instruction and/or do reading activities with the students, how often do you teach students strategies for decoding sounds and words? (Teachers were
asked to select one of the following options: Every day or almost every day, once or twice a week, once or twice a month, never or almost never)

These questions were selected based upon findings from previous research which indicates the importance of teachers’ use of particular instructional strategies and activities on reading. Specifically, previous research highlights the importance of teaching using small groups in addition to regular whole-class instruction to support students to improve their literacy skills (O’Sullivan, Canning, Siegel, & Oliveri, 2009). For example, findings from a study conducted by Slavin, Cheung, Groff and Lake (2008) suggest evidence of moderate effectiveness of small group on learning. Findings from another study conducted by Lesaux and Siegel (2003) similarly point to the importance of teaching students in small groups. This study suggests that such instruction should focus on explicit instruction of key factors associated with reading such as teaching students decoding and phonological awareness. Several studies (Siegel 1993; Stanovich 1988; Stanovich & Siegel 1994) also highlight the importance of decoding as well as the automaticity with which students can decode words as key factors underlying reading and suggest that phonological processing, which involves decoding, is one of the most significant underlying cognitive processes used in reading acquisition. Findings from the abovementioned research studies and research reviews were used as pointers in the selection of variables potentially underlying differences in examinees’ item responding.

2.3.5. Comparison Groups and Samples

This study was conducted using data administered to representative samples of examinees from four countries: Chinese Taipei, Hong Kong, Kuwait and Qatar. These examinees took the test in non-Indo European languages. Specifically, examinees from the first two countries took the test in Mandarin whereas examinees from the latter countries took the test in Arabic. These language groups were selected for two reasons. First, findings from previous
studies suggest greater degrees of DIF for group comparisons that are from different language groups (Ercikan & Koh, 2005; Grisay, Gonzalez, & Monseur, 2009; Grisay & Monseur, 2007; Oliveri & von Davier, 2011). Second, examinees from these countries were also expected to differ with respect to instructional methods and curricula. For example, as outlined in the PIRLS 2006 Encyclopedia (Kennedy, Mullis, Martin, & Trong, 2007), as a way of learning to read, the curriculum from Qatar emphasized reading the Quran whereas the Chinese Taipei and Hong Kong curricula emphasized having exposure to a variety of texts from multicultural contexts (Kennedy et al., 2007). Further, whereas the curricula from Kuwait and Qatar emphasized learning to decode accurately and read fluently in Grades 3 to 5, the curricula from Hong Kong and Chinese Taipei emphasized understanding, organizing information, discussing and criticizing reading material as well as forming literate citizens to think independently (Kennedy et al., 2007). Given the focus on instructional methods as potential sources of DIF in this study, these country comparisons presented an important opportunity for this investigation.

Descriptive statistics for each of the four countries (Chinese Taipei, Hong Kong, Kuwait and Qatar) are given in Table 2.4. The first two columns list the language group and country examined, the third column describes sample sizes, columns 4 and 5 are the number of boys and girls that took the Reader booklet, and columns 6 and 7 are each country’s combined (informational and literary reading-comprehension) mean scale score and mean total test raw scores, respectively. For each country, sample sizes listed in Table 2.4 represent the entire group of examinees who took the PIRLS Reader. Sampling weights were not used in this study. These weights are typically used in ELSAs to accommodate non-proportional representation (over- or under-representation) of some groups in the obtained sample compared to their representation in the population (Rutkowski et al., 2010). The weights are obtained for adjusting the total sample within a country and are not designed to adjust for sampling non-representativeness for samples at the booklet level. As this study focused
only on one booklet (i.e., the Reader), sampling weights were not used. In addition, sampling weights are essential when group level performances are compared; however, they are not necessary when examining DIF which focuses on response patterns (rather than group level results).

Table 2.4

**Reader Booklet Descriptive Statistics**

<table>
<thead>
<tr>
<th>Language</th>
<th>Country</th>
<th>Sample Size</th>
<th>Boys</th>
<th>Girls</th>
<th>Mean Scale Score</th>
<th>Mean Raw Score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>Taipei</td>
<td>914</td>
<td>479</td>
<td>435</td>
<td>535 (2.0)</td>
<td>19.77 (5.91)</td>
</tr>
<tr>
<td></td>
<td>Hong Kong</td>
<td>949</td>
<td>481</td>
<td>468</td>
<td>564 (2.4)</td>
<td>21.37 (5.42)</td>
</tr>
<tr>
<td></td>
<td>Total/Average</td>
<td>1,863</td>
<td>960</td>
<td>903</td>
<td>550</td>
<td>19.84</td>
</tr>
<tr>
<td>Arabic</td>
<td>Kuwait</td>
<td>757</td>
<td>345</td>
<td>412</td>
<td>330 (4.2)</td>
<td>7.1 (5.11)</td>
</tr>
<tr>
<td></td>
<td>Qatar</td>
<td>1,315</td>
<td>659</td>
<td>656</td>
<td>353 (1.1)</td>
<td>6.59 (4.32)</td>
</tr>
<tr>
<td></td>
<td>Total/Average</td>
<td>2,072</td>
<td>1,004</td>
<td>1,068</td>
<td>342</td>
<td>6.26</td>
</tr>
</tbody>
</table>

2.4. Analyses

This study investigated the following three research questions: (1) What is the degree of WGH for the four selected countries with respect to the underlying structure of the PIRLS 2006 Reader? (2) To what extent do manifest DIF detection and latent class DIF methods result in the identification of the same DIF items? and (3) What are the sources of incomparability underlying differentially functioning items? Analyses to investigate these three research questions involved various steps and procedures which will be described in more detail in the next subsections.

The following subsections will specifically describe the particular methods used to investigate unidimensionality and group homogeneity as well as DIF and its sources. The dimensionality of the PIRLS 2006 Reader was examined to investigate the number of factors underlying the structure of the
measure. As the measure assessed only one construct (i.e., reading), test of a one factor solution was investigated using confirmatory factor analysis (CFA). The degree of group homogeneity was examined using the discrete mixture distribution (DMD) model. Two methods were used to detect manifest DIF: the Logistic Regression (LR) and IRT based L-H DIF detection models. Latent class DIF was investigated using the DMD model. Sources of DIF were examined by systematically grouping DIF items according to three principles: (1) purpose for reading, (2) reading comprehension process, and (3) cognitive complexity of items. This was done to determine the potential relationship between DIF and the types of reading processes required by the test items. The composition of LCs in relation to selected student- and teacher-level variables was investigated using two methods: the discriminant function analysis and the multinomial logistic regression models. Moreover, due to nesting of the sample within intact classrooms, a multilevel multinomial logistic regression model was used.

2.4.1. Analysis of Unidimensionality

The assumption of unidimensionality, implies that there is only one dimension needed to account for item responses in the measure. Often the assumption of strict unidimensionality is not met in educational achievement tests given the diverse curricular content knowledge covered by such tests. Instead, the presence of a single ‘dominant’ factor is considered sufficient for meeting unidimensional assumptions required by manifest DIF methods such as those based on IRT (Hambleton et al., 1991).

The unidimensionality of the PIRLS 2006 Reader was investigated using CFA using the Mplus software (Muthén & Muthén, 1998-2007; Muthén & Muthén, 2008). CFA was conducted using the robust weighted least squares (WLSMV) estimator, which was selected because the factor indicators were categorical (Brown, 2006). CFA is a statistical method used to examine the number of underlying dimensions (i.e., continuous factors or latent variables) contained in a set of observed
variables (e.g., test items). As the PIRLS 2006 Reader was developed to measure one primary construct (e.g., reading), fit of a one-factor solution was investigated. The following three goodness-of-fit indices were used to examine model fit: (1) root mean square error of approximation (RMSEA; Steiger, 1990), (2) Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) and (3) the weighted root-mean-square residual (WRMR; Yu, 2002).

The RMSEA is a standardized index which assesses the extent to which a model fits a specific data set. It is the average discrepancy between correlations observed in the input matrix and those predicted by the model. It includes a penalty function for models that use too many parameters when a more parsimonious model could fit the data equally well (Brown, 2006). A value of zero indicates no discrepancy and, therefore, a perfect fit of the model to the data. On the other hand, models with RMSEA > 0.6 have poor fit and should be rejected (Brown, 2006).

The TLI is a comparative fit index, wherein the fit of a user-specified model is tested against a baseline model. The baseline model is referred to as a ‘null’ model in which no relationships among variables are specified. The TLI compensates for the effect of model complexity by including a penalty function for estimating parameters that do not significantly contribute to the estimated model. Values range from 0.0 to 1.0, wherein values above .90 are considered acceptable fit (Brown, 2006).

The WRMR measures the (weighted) average differences between the sample and estimated population variances and covariances and has been recently proposed as suitable for models using ordinal item responses (Muthén & Muthén, 1998-2007). A cutoff value below 1.0 indicates good fit (Yu, 2002).

2.4.2. Analysis of DIF

The next step entailed assessing DIF using two approaches: a manifest and a latent class DIF approach. Analyses of manifest DIF sought to investigate whether individuals of equal ability showed differences in item performance. The following four pairwise country comparisons were examined:
Hong Kong versus Kuwait, Taipei versus Kuwait, Hong Kong versus Qatar and Taipei versus Qatar. Two manifest DIF detection methods were used: the IRT based L-H, and LR. Latent class DIF analyses were conducted to examine the presence and number of LCs in the data as well as particular items functioning differentially across each of the identified LCs. These analyses were conducted using the DMD model. Once LCs were identified, the correspondence between items identified as DIF using the DMD model and those detected using the IRT based L-H and LR methods was investigated using examinee LC membership as a grouping variable.

2.4.3. Detection of DIF Using a Manifest Approach

Analysis of manifest DIF was conducted using the IRT based L-H and LR methods to increase accuracy in DIF detection (Ercikan et al., 2004). These two methods were selected because they allow for estimation of DIF for dichotomous and polytomous items as well as uniform and non-uniform DIF (Clauser & Mazor, 1998). The effects of language on performance may differ along the ability scale. Therefore, methods that are designed to detect both uniform and non-uniform DIF were used (Oliveri & Ercikan, 2011).

The IRT based L-H method involves computation of the difference in deciles between the predicted and observed probability of correctly responding to an item or of obtaining the maximum score ($P_{diff}$). Based on this difference, a chi-square ($\chi^2$) statistic is obtained and converted into a Z-statistic. DIF is identified as being in favour of or against a comparison group according to a statistical significance level as well as an effect size ($p_{diff}$). Level 2, moderate DIF, was identified if $|Z| \geq 2.58$ and $|p_{diff}| < 0.10$ and Level 3, large DIF, was identified if $|Z| \geq 2.58$ and $|p_{diff}| \geq 0.10$. MC items were calibrated using a 3PL model and CR items were calibrated using the two-parameter partial credit model (2PPC; Yen, 1993), which is equivalent to Muraki’s (1992) Generalized Partial Credit Model. To estimate IRT parameters, items were jointly calibrated using PARDUX (CTB/McGraw-Hill, 1991).
Second, the LR method (Swaminathan & Rogers, 1990) was used to analyze DIF with dichotomous items and Ordinal Logistic Regression (OLR; Zumbo, 1999) was used to analyze DIF for polytomous items. The LR method is based on a logistic regression equation, and involves the regression of the total (or rest) score, group membership (e.g., language, gender or ethnic background), and the interaction between total (or rest) score and group membership. The total score is the sum of examinees’ responses to test items. The rest score is the total score minus the studied item. The regression equation is conducted in a step-wise fashion resulting in a $\chi^2$ based test of significance and an effect size ($\text{Nagelkerke } R^2$) obtained for each step. Differences between $R^2$ values with the addition of each new step in the model provide a measure of variance accounted for by the added variable (Zumbo, 1999). Using this method, items were identified as DIF if the $\chi^2$ value from the grouping variable is statistically significant at $p < 0.01$, and items were considered to have negligible (or A-level) DIF if $R^2 < 0.035$, moderate DIF if $0.035 \leq R^2 \leq 0.07$, and large DIF if $R^2 > 0.07$ (Jodoin & Gierl, 2001).

### 2.4.4. Identification of DIF Using a Latent Class Approach

The identification of latent class DIF was conducted using the DMD model (von Davier & Yamamoto, 2004). The DMD model allows for estimating models wherein grouping variables are not known a priori but instead are estimated based upon examinees’ item responses. It also allows for the estimation of binary (Rost, 1990) and polytomous Rasch models (von Davier & Rost, 1995) as well as 2PL IRT models (Mislevy & Verhelst, 1990; Kelderman & Macready, 1990). The DMD model was estimated using the multidimensional discrete latent trait models (mdltm) software (von Davier, 2005b).

Analyses of latent class DIF were conducted in three steps. The first step involved examining sample homogeneity and identifying the best fitting model. To this end, various models (starting with one LC) were estimated; the number of LCs was increased one at a time, until the best fitting model
was identified. The second step involved examining model-data fit associated with the resulting LCs using the Bayesian Information Criterion (BIC; Schwarz, 1978). According to Schwarz (1978), the BIC is a minimum discrepancy index that penalizes the addition of model parameters with a term that depends on the sample size; in this way, the index is assumed to ensure that the model with minimal BIC values represents the relatively best-fitting model. The penalty term of the BIC takes the sample size into account so that over-parameterization due to large samples is less of a problem for the BIC as compared to the use of the Akaike information criterion (AIC; Akaike, 1974). The BIC was used based upon findings from previous simulation studies (Lubke & Spies, 2008; Lubke & Neale, 2008) that suggest that the BIC leads to more accurate estimates of fit as compared to other criteria (e.g., AIC or sample adjusted BIC).

The last step involved comparing item difficulty parameters for the potentially resulting LCs to identify differentially functioning items. The following differences in $b$ parameters were used to classify DIF items: small DIF ($0.40 \leq \Delta b < 0.80$), moderate DIF ($0.80 \leq \Delta b < 1.20$), and large DIF ($\Delta b \geq 1.20$) (Samuelsen, 2005). In the DMD model, the difficulty parameters can be specified to be linked in the scaling process; thus, enabling difficulty parameters to be comparable across LCs (von Davier, 2008).

2.4.5. Investigation of Sources of DIF

Once DIF was identified using a manifest or a latent class DIF approach, the next step was to examine potential sources of DIF. Substantive analyses of DIF were conducted to determine if there were any patterns in DIF items with regards to the following three substantive aspects used to group items: (1) reading comprehension process, (2) reading purpose and (3) cognitive complexity of items. These three aspects were described in the assessment framework (section 2.3.2.).
2.4.6. Investigation of the Composition of the Resulting Latent Classes

Analyses of the composition of resulting LCs involved three steps. First, average plausible values provided in the PIRLS dataset were used to investigate average student performance in relation to examinees’ country. Second, the proportion of boys and girls and Mandarin- versus Arabic-speaking examinees associated with each LC was investigated. These manifest variables were selected because they have been typically used for grouping examinees when conducting manifest DIF analyses. Third, the multinomial logistic regression and descriptive discriminant function analysis (Tabachnick & Fidell, 2007) models were run to examine LC composition. As the DV (LC membership) was an unordered categorical variable, a multinomial model was used. Three LC comparisons were conducted: LC1 versus LC2, LC1 versus LC3 and LC2 versus LC3. Moreover, a multinomial multilevel logistic regression model was run due to nesting of students within classrooms; results of these analyses are presented in Appendix B.

2.5. Results

2.5.1. Results of Unidimensionality Analyses

Analyses of the number of dimensions underlying the test data structure were conducted using CFA with a WLSMV estimator. Fit was evaluated using three goodness-of-fit criteria: the TLI, the RMSEA and the WRMR. These results are presented in Table 2.5. Results of the TLI and RMSEA fit indices suggest acceptable fit of a one factor model for each of the four countries.
Table 2.5

*Goodness-of-Fit Criteria for Evaluating Unidimensionality*

<table>
<thead>
<tr>
<th></th>
<th>Qatar</th>
<th>Hong Kong</th>
<th>Chinese Taipei</th>
<th>Kuwait</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSEA</td>
<td>.019</td>
<td>.028</td>
<td>.029</td>
<td>.037</td>
</tr>
<tr>
<td>TLI</td>
<td>.986</td>
<td>.960</td>
<td>.972</td>
<td>.964</td>
</tr>
<tr>
<td>WRMR¹</td>
<td>.949</td>
<td>1.065</td>
<td>1.049</td>
<td>1.102</td>
</tr>
</tbody>
</table>

*Note:* RMSEA = Root Mean Square Error of Approximation  
TLI = Tucker-Lewis Index  
WRMR = Weighted Root-Mean-Square Residual

2.5.2. Results of Manifest DIF Analyses

Analyses to investigate DIF were conducted using the following pairwise country comparisons: Hong Kong versus Kuwait, Taipei versus Kuwait, Hong Kong versus Qatar and Taipei versus Qatar. Results of these analyses using the IRT and LR methods of DIF detection are summarized in Table 2.6. LR results are based on the rest score (the total score minus the studied item) (Junker & Sijtsma, 2000). The first column lists the items detected as functioning differentially across country comparisons. The remaining columns summarize DIF detection results for each pairwise comparison. Items detected as DIF using only the IRT method are noted using the following notation (√IRT), whereas DIF items detected using the IRT and LR methods are noted as (√IRT/LR); there were no items detected as DIF only by the LR method. To illustrate, columns 2 and 3 indicate the comparison between Hong Kong versus Kuwait. In total, there was one item favouring Hong Kong, which was detected by LR and IRT indicated by (√IRT/LR) (see column 2). As shown in column 3, there were 4 items favouring Kuwait, all of which were detected using IRT only (√IRT). All detected items had WRMR values for all countries except Qatar were above the 1.0 suggested cutoff value for good fit. As suggested by Muthén (2010) the WRMR remains an experimental fit statistic. If the fit of the other statistics reveal acceptable fit, the higher WRMR values should not be of concern.

¹ WRMR values for all countries except Qatar were above the 1.0 suggested cutoff value for good fit. As suggested by Muthén (2010) the WRMR remains an experimental fit statistic. If the fit of the other statistics reveal acceptable fit, the higher WRMR values should not be of concern.
It is of interest to note that across all comparisons, LR detected a lower number of DIF items as compared to IRT. Differences in the number of DIF items across the IRT based L-H and LR methods
may partially be explained because these methods differ in two ways. First, the probability of correct response (or obtaining the maximum score) in these methods is estimated differently as is the way in which examinees are matched. In LR, the probability of correct response is the number of people who obtain the maximum score divided by the total number of examinees in the matched ability group. In IRT, the probability of obtaining the maximum score is the value of the item response function for range of theta levels estimated using maximum likelihood functions which jointly estimate examinee and item parameters. Moreover, in LR the matching criterion is the total number correct score and in IRT the criterion is based on theta estimated jointly with item parameters using marginal maximum likelihood estimation techniques (Oliveri & Ercikan, 2011).

2.5.3. Results of Latent Class DIF Analyses

The first step in the analysis of latent class DIF was to evaluate the homogeneity of the sample. These analyses were conducted by running the data starting with one LC and gradually increasing the number of LCs by one; models ranging from one to four LCs were tested. The second step was comparing BIC results across estimated models; recall that the lowest comparative value in the BIC indicates improved fit estimates (Lubke & Neale, 2008; Lubke & Spies, 2008). Results of these analyses are shown in Table 2.7. Column one indicates the number of LCs evaluated (LC1 through LC4), column 2 indicates the number of parameters associated with each of the LCs evaluated, column 3 shows the BIC statistic used for evaluating model-data fit and columns 4 through 7 indicate class proportions associated with each of the examined solutions. As shown in Table 2.7 (see column 3) improved model-data fit estimates were obtained when three LCs were specified as it resulted in the lowest BIC values. As shown in columns 4 through 6, the smallest LC had 27% of the sample and the largest LC had 39% of the sample.
### Table 2.7

**Fit Statistics for the PIRLS 2006 Reader**

<table>
<thead>
<tr>
<th>No. of LCs</th>
<th>P</th>
<th>BIC</th>
<th>LC1</th>
<th>LC2</th>
<th>LC3</th>
<th>LC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59</td>
<td>106785.13</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>117</td>
<td>106630.81</td>
<td>.68</td>
<td>.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>176</td>
<td>106594.11</td>
<td>.27</td>
<td>.39</td>
<td>.33</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>235</td>
<td>106859.25</td>
<td>.30</td>
<td>.14</td>
<td>.39</td>
<td>.17</td>
</tr>
</tbody>
</table>

*Note: P = # of parameters estimated*

BIC = Bayesian Information Criteria

LC = Latent Class

Resulting LCs differed in ability distributions. Specifically, LC3 was the highest performing class ($M=0.817; SD=0.434$) and consisted of 1,307 examinees. The next highest performing class was LC1 ($M=.081; SD=.601$) and had 1,081 examinees. The lowest achieving class was LC2 ($M=-0.89; SD=0.456$) and had a total of 1,547 examinees.

The third step in the analysis of latent class DIF was to examine item difficulty parameters associated with each of the three resulting LCs. Findings revealed that there were 18 out of 26 (69%) items displaying DIF; that is, differences in $b$ parameters of at least $\Delta b \geq 0.40$ across LCs. Criteria described in Samuelsen (2005) were used to classify items according to their effect size. As shown in Table 2.8, findings revealed that one item had small DIF ($0.40 \leq \Delta b < 0.80$), five items had moderate DIF ($0.80 \leq \Delta b < 1.20$) and twelve items had large DIF ($\Delta b \geq 1.20$).


Table 2.8

Effect Size of DIF Identified Using a Latent Class DIF Approach

<table>
<thead>
<tr>
<th>DIF Favouring</th>
<th>DIF Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>LC 1</td>
<td></td>
</tr>
<tr>
<td>LC 2</td>
<td>0</td>
</tr>
<tr>
<td>LC 3</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: LC=Latent Class

To compare results of latent variable and manifest DIF methods, further analyses were conducted using LR and IRT methods using LC membership as a grouping variable. The rest score was used for conducting LR DIF analyses. Results are summarized in Table 2.9. The first column specifies the item identified as DIF using the latent class DIF approach, the second to fourth columns show the method used to analyze DIF, and the fifth to seventh columns indicate the correspondence between the latent class DIF approach and whether the item was detected as differentially functioning by 0, 1 or 2 manifest methods, respectively. Results indicated that among the 18 items identified as DIF using the latent variable mixture model, IRT identified 12 (67%) of the same items; two (17%) had a large effect size and ten (83%) had a moderate effect size. In addition, IRT identified two other items that had not been identified by the latent class DIF model. The LR method detected eight (44%) of the items that had been identified by the latent class DIF model. Two (25%) of the items had a large effect size and six (75%) had a moderate effect size.
Table 2.9

Correspondence between Items Identified by Latent Class and Manifest DIF Approaches

<table>
<thead>
<tr>
<th>ITEM</th>
<th>DIF Method</th>
<th>Correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LC</td>
<td>IRT</td>
</tr>
<tr>
<td>U1</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>U2</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U3</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>U4</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U5</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U6</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U7</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U8</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U10</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>U12</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S13</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Total # of items</td>
<td>18</td>
<td>12</td>
</tr>
</tbody>
</table>

Note: LC DIF=Latent Class DIF
IRT=Item Response Theory
LR=Logistic Regression

* Items detected as having large DIF

Results of these analyses indicated that the clustering of examinees using LC membership, as compared to a manifest grouping variable, resulted in the identification of a higher number of DIF
items. Recall that when using test language as a manifest grouping variable, LR and IRT detected two and seven items as DIF, respectively. On the other hand, when using LC membership as a grouping variable, LR identified eight items and IRT identified 14 items as DIF.

2.5.4. Results of Substantive Analyses

To identify sources of DIF, substantive analyses were conducted. These analyses consisted of systematically grouping items identified as functioning differentially using either a manifest or a latent class DIF approach according to the following three criteria: (1) reading purpose, (2) reading comprehension process, and (3) cognitive complexity of items.

Results related to the systematic grouping of manifest DIF items according to purposes of reading, reading comprehension process and cognitive complexity of items are summarized in Table 2.10, see column 1. Columns 2 through 9 show results of pairwise country comparisons, starting with Hong Kong versus Kuwait, Hong Kong versus Qatar, Chinese Taipei versus Kuwait and Chinese Taipei versus Qatar. Rows 3 to 5 list whether the DIF items measured students’ ability to read for pleasure (also referred to as reading for literary experience) or for information. Rows 6 through 10 list the comprehension process measured by the item (i.e., making inferences, retrieving, interpreting or examining information). Rows 11 through 13 indicate the cognitive complexity assessed by the item (i.e., low versus high cognitive complexity). To illustrate, in the Hong Kong versus Kuwait comparison, there were five DIF items. Four items favoured Kuwait and one item favoured Hong Kong. Among the items favouring Kuwait, two assessed making inferences, one item assessed retrieving information and the other assessed interpreting information. The item favouring Hong Kong assessed interpreting information.

As shown in Table 2.10, there were split results for the comparison groups in relation to reading comprehension process as well as cognitive complexity of items. For instance, in the comparison of Chinese Taipei versus Kuwait, among the seven items favouring Kuwait, three (43%)
assessed making inferences, one (14%) retrieving information, two (29%) interpreting information and one examining information. Items were thus also split in relation to cognitive complexity with four items examining low cognitive skills and three items measuring high cognitive skills. Among the two items favouring Taipei, one assessed making inferences and one item measured interpreting information; these items were evenly split in relation to the measurement of cognitive skills. Similarly split results were found in relation to cognitive complexity of items associated with comparison groups. To illustrate, in the Hong Kong versus Qatar comparison, among the three items favouring Qatar, two (66%) assessed low cognitive complexity skills and one (33%) assessed high cognitive complexity. The item favouring Hong Kong assessed low cognitive complexity. More consistent patterns were seen, however, in relation to reading purpose. To exemplify, in the Chinese Taipei versus Qatar comparison, 100% of items favouring Qatar assessed reading for pleasure and 100% of items favouring Chinese Taipei assessed reading for information.
Table 2.10

**Grouping of Manifest DIF Items**

<table>
<thead>
<tr>
<th>Purpose</th>
<th>H vs. KU</th>
<th>H vs. QA</th>
<th>T vs. KU</th>
<th>T vs. QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleasure</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Information</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Comp. Process</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inference</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Retrieve</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Interpret</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Examine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cog. Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note: H=Hong Kong, KU = Kuwait, QA = Qatar, T= Chinese Taipei
Comp. Process = Reading Comprehension Process
Cog. Complexity = Cognitive Complexity

Results related to the systematic grouping of latent class DIF items according to purposes of reading, reading comprehension process and cognitive complexity of items are summarized in Table 2.11. The information presented in this table is organized in the same way as Table 2.10. As shown in Table 2.11, LCs differed in relation to each other according to purposes of reading, reading comprehension process and cognitive complexity of items suggesting that examinees from each of the LCs had distinct patterns of item responding. To illustrate, in relation to reading purpose, six out
of seven items (86%) favouring LC1 assessed reading for literary experience. Seven out of nine items (78%) favouring LC3 assessed reading for information. Five out of seven (71%) items favouring LC1 were associated with the reading comprehension process of making inferences, 100% of items favouring LC2 were related to retrieving information and seven out of nine items (78%) favouring LC3 were related to interpreting information. In relation to cognitive complexity, six out of seven (86%) items favouring LC1 assessed lower-order cognitive complexity; 100% of items favouring LC2 also assessed lower-order cognitive complexity and eight out of nine items (89%) favouring LC3 assessed higher-order cognitive complexity.
Table 2.11

*Grouping of Latent Class DIF Items*

<table>
<thead>
<tr>
<th>Latent Class</th>
<th>Pro LC1</th>
<th>Pro LC2</th>
<th>Pro LC3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pleasure</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Information</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td><strong>Comp. Process</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inference</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Retrieve</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Interpret</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Examine</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Cog. Complexity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

*Note:* Comp. Process = Reading Comprehension Process  
Cog. Complexity = Cognitive Complexity  
LC = Latent Class

2.5.5. **Results of Analyses of Latent Class Composition**

The composition of the resulting LCs was examined in relation to groupings related to country, gender and language of test. These results are presented next.

Results of the analyses examining to what extent LCs corresponded to country groupings are shown in Table 2.12. For these analyses, average plausible values (i.e., the average of the five plausible values provided in the PIRLS data set) were used. In Table 2.12, column 1 identifies the
LCs, columns 2 to 5 are average plausible values associated with each of the participating countries and column 6 is the average plausible value of each LC. As shown in Table 2.12, LC1 and LC2 respectively had middle- and lowest performing students of all countries except Hong Kong. LC3 consisted of top performing students across all countries.

Table 2.12

*Average Plausible Value Scores Associated with Examinees’ Country of Residence and LC*

<table>
<thead>
<tr>
<th></th>
<th>Taipei</th>
<th>Hong Kong</th>
<th>Kuwait</th>
<th>Qatar</th>
<th>Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC1</td>
<td>516.88</td>
<td>509.41</td>
<td>431.66</td>
<td>445.93</td>
<td>474.7</td>
</tr>
<tr>
<td>LC2</td>
<td>419.71</td>
<td>545.48</td>
<td>285.84</td>
<td>305.71</td>
<td>317.2</td>
</tr>
<tr>
<td>LC3</td>
<td>576.91</td>
<td>579.22</td>
<td>459.71</td>
<td>499.71</td>
<td>571.27</td>
</tr>
</tbody>
</table>

*Note: LC=Latent Class*

Results of the analyses examining to what extent LCs corresponded to gender groupings revealed that each LC had varying proportions of boys and girls. To illustrate, there were 29% of boys in LC1, 37% in LC2 and 34% in LC3 and there were 25% of girls in LC1, 42% in LC2 and 33% in LC3. Similarly, results of LC composition related to language of the test revealed that each LC had varying proportions of Mandarin- and Arabic-speaking examinees. For instance, there were 27% of Mandarin-speaking examinees in LC1, 8% in LC2 and 65% in LC3 and there were 28% of Arabic-speakers in LC1, 68% in LC2 and 4% were in LC3.

Further analyses were conducted to examine the composition of resulting LCs in relation to selected student- and teacher-level variables. These analyses were conducted using two different models: a descriptive discriminant function analysis and a multinomial logistic regression analysis model. These models were run using LC membership as a dependent variable (DV). These two methods are complementary to each other. That is, whereas descriptive discriminant function analyses yield descriptive results indicating the location of group centroids; logistic regression
analyses provide statistical significance values for each predictor variable. Moreover, whereas multiple groups can be analyzed simultaneously when using discriminant function analyses, groups are examined pairwise with regression analysis. A third method, multilevel multinomial logistic regression analyses, was also used due to nesting of the data associated with sampling of students within classrooms. These results are presented in Appendix B.

Results of the multinomial logistic regression analyses are summarized in Table 2.13; variables that were found to be significant are in bold. Results indicated that average plausible value was a significant predictor across LCs. The odds ranged from .976 to 1.029, which indicated that an increase in plausible values led to substantially higher odds of being in a higher performing LC. Above and beyond average plausible values, the following variables were found to be significant: gender (LC1 vs. LC2) and country (Taipei versus Qatar, Hong Kong versus Qatar, and Kuwait versus Qatar). Gender was significant for LC1 versus LC2 (odds ratio=1.59) suggesting that being a female increased the odds of being in the higher achieving LC1 versus LC2 by 59%. The comparison between Taipei versus Qatar was significant for LC1 versus LC2 and LC2 versus LC3 (odds ratio=0.609 and 0.344, respectively) indicating that being from Taipei led to slightly higher odds of being in a higher performing LC. The comparison between Hong Kong versus Qatar was significant for LC1 versus LC2 and LC1 versus LC3 (odds ratio=5.462 and 2.88, respectively), which indicated that being from Hong Kong led to substantially higher odds of being in a higher performing LC. Finally, the comparison between Kuwait versus Qatar was significant for LC1 versus LC2 (odds ratio=0.411) suggesting that being from Kuwait led to slightly higher odds of being in the higher performing LC (LC1).
Table 2.13

*Results of Multinomial Logistic Regression Analyses*

<table>
<thead>
<tr>
<th></th>
<th>LC 1 vs. 2</th>
<th>Odds</th>
<th>LC 1 vs. 3</th>
<th>Odds</th>
<th>LC 2 vs. 3</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plausible Values</td>
<td>&lt;.01***</td>
<td>.976</td>
<td>&lt;.01***</td>
<td>1.029</td>
<td>&lt;.01***</td>
<td>.977</td>
</tr>
<tr>
<td>Age</td>
<td>.934</td>
<td>NS</td>
<td>.051</td>
<td>NS</td>
<td>.384</td>
<td>NS</td>
</tr>
<tr>
<td>Gender</td>
<td>&lt;.01***</td>
<td>1.590</td>
<td>.811</td>
<td>NS</td>
<td>.473</td>
<td>NS</td>
</tr>
<tr>
<td>Taipei vs. Qatar</td>
<td>&lt;.01***</td>
<td>.609</td>
<td>.785</td>
<td>NS</td>
<td>&lt;.01***</td>
<td>.344</td>
</tr>
<tr>
<td>Hong Kong vs. Qatar</td>
<td>&lt;.01***</td>
<td>5.462</td>
<td>&lt;.01***</td>
<td>2.880</td>
<td>.373</td>
<td>NS</td>
</tr>
<tr>
<td>Kuwait vs. Qatar</td>
<td>&lt;.01***</td>
<td>.411</td>
<td>.166</td>
<td>NS</td>
<td>.491</td>
<td>NS</td>
</tr>
<tr>
<td>Small Group Teaching</td>
<td>.953</td>
<td>NS</td>
<td>.146</td>
<td>NS</td>
<td>.927</td>
<td>NS</td>
</tr>
<tr>
<td>Same Ability Groups</td>
<td>.295</td>
<td>NS</td>
<td>.230</td>
<td>NS</td>
<td>.337</td>
<td>NS</td>
</tr>
<tr>
<td>Decoding</td>
<td>.881</td>
<td>NS</td>
<td>.691</td>
<td>NS</td>
<td>.894</td>
<td>NS</td>
</tr>
</tbody>
</table>

*Note: LC=Latent Class*

*** p<.01
NS= Not Significant

Results of the discriminant function analysis indicated that two discriminant functions were statistically significant. As indicated by the Wilk’s lambda ‘peel off’ test (Tabachnick & Fidell, 2007), after the first function ($\chi^2=3009.553$) was removed, the test of discriminant function 2 with $\chi^2=138.765$ was also statistically significant at $\alpha=.05$. The first discriminant function accounted for 97.3% of the explained variance and the second function explained the remaining 2.7% of the between-group variance. As shown in Figure 2.1, the first discriminant function separates LC2 from LC3 and the second function separates LC1 from LC2 and LC3.
The following two predictors were found to correlate with the first function, which explained 97.3% of the variance: average student performance (measured by average plausible value) \((r=.982)\) and teaching same ability groups \((r=.334)\). The following predictors correlated with the second function, which explained 2.7% of the between group variance: the comparisons between Taipei versus Qatar \((r=.526)\) and Hong Kong versus Qatar \((r=-.675)\). Finding significance of average plausible value was consistent with multinomial logistic regression results as were results comparing Taipei versus Qatar and Hong Kong versus Qatar; however these country comparisons explained only 2.7% of between group variance. Results also indicated that the following two variables were not highly correlated with either of the two discriminant functions: teaching strategies for decoding \((r=.211)\) and percentage time working with small groups \((r=.292)\). These findings were consistent with results from the multinomial logistic regression. The other variables (gender, age and country: Kuwait versus Qatar) also had correlations below 0.3 and were not interpreted.
Table 2.14

*Discriminant Function Analysis Structure Matrix*

<table>
<thead>
<tr>
<th></th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Average plausible value</td>
<td>.982</td>
</tr>
<tr>
<td>Percentage time working with small groups</td>
<td>-.217</td>
</tr>
<tr>
<td>Teach strategies for decoding</td>
<td>-.083</td>
</tr>
<tr>
<td>Teaching same ability groups</td>
<td>.334</td>
</tr>
<tr>
<td>Taipei versus Qatar</td>
<td>.400</td>
</tr>
<tr>
<td>Hong Kong versus Qatar</td>
<td>.469</td>
</tr>
<tr>
<td>Kuwait versus Qatar</td>
<td>.110</td>
</tr>
<tr>
<td>Age</td>
<td>.075</td>
</tr>
<tr>
<td>Gender</td>
<td>.025</td>
</tr>
</tbody>
</table>

2.6. Discussion

This study investigated the assumption of WGH associated with manifest DIF detection methods. The use of a latent class DIF approach using the DMD model allowed for investigation of sample homogeneity. DIF and its sources were examined using two different approaches: a manifest and a latent class DIF approach. This investigation was conducted using data from the PIRLS 2006 Reader administered to examinees from four countries: Hong Kong, Chinese Taipei, Kuwait and Qatar.

Different results were obtained from the use of a manifest versus a latent class DIF approach in relation to which items were found to function differentially, the degree of DIF, and the sources of the identified DIF. First, the use of a manifest DIF approach led to detecting a lower number of DIF
items with smaller effect sizes. For example, the use of a manifest approach detected only two DIF
items with large effect sizes whereas 12 items with large effect sizes were identified when a latent
variable approach was used. This result suggests that the grouping of examinees using a manifest
characteristic (e.g., country) may have led to under detecting a proportion of DIF items as well as
their magnitude. This study’s design did not allow for systematically investigating the effects of
heterogeneity on manifest DIF detection rates. This type of examination requires a simulation study
design and has been more thoroughly investigated in study two.

Investigation of LC membership revealed that the primary factor differentiating LCs was
reading achievement (as measured by average plausible values). These findings were consistent
across methods (discriminant function and multinomial logistic regression analyses). This finding
may not be surprising given that the test is intended to measure performance in reading. However,
investigation of substantive aspects used to group DIF items provided additional information not only
in regards to how these LCs differed in overall achievement but also in relation to how these LCs
differed in their response patterns. For example, a large proportion (86%) of items favouring LC1
were related to reading for pleasure whereas a large proportion (78%) of items favouring LC3 were
associated with reading for information, indicating particular types of reading abilities. A similar
pattern was observed across LC1 and LC3 in relation to reading comprehension skills, wherein a
large proportion (86%) of items favouring LC1 measured retrieving information and making
inferences (lower-order reading comprehension skills) and 89% of items favouring LC3 assessed
examining, evaluating and interpreting text (higher-order reading comprehension skills).

The use of a latent class DIF approach, as compared to the use of a manifest DIF detection
approach, yielded stronger associations in relation to substantive aspects used to group DIF items.
These results suggest that a latent class approach may lead to clearer results regarding examinees’
response patterns. It is possible that conducting linguistic reviews of items or examining cultural
differences in how examinees from each of the examined countries interpreted items may have provided further information about sources of manifest DIF. These issues were not examined in this study and should be the subject of future research.

Analyses of LC composition in relation to classroom and teacher related variables (e.g., teaching in small groups, decoding) yielded less conclusive results. This investigation was conducted using only three variables from the teacher background questionnaire, it is possible that examining a larger number and a wider array of variables related to teacher and classroom instruction may have led to more conclusive findings regarding which variables had a strong effect in differentiating across LCs. Results of LC composition analyses in relation to manifest variables (e.g., gender and country) used to group examinees when conducting DIF analyses using a manifest DIF detection method were either not significant, or if they were significant, they explained a small amount of between-group variance. These findings suggest that manifest characteristics typically used for grouping examinees when conducting manifest DIF analyses may not be well aligned with the strongest factors underlying students’ differences in item responses. This finding suggests that alternative variables should be used to investigate DIF. However, this suggestion does not address some of the purposes for which DIF is investigated: policy-and legal-decisions that are conducted based on manifest variables such as ethnic background or gender. DeMars and Lau (2011) suggest that “For policy decisions and legal documentation, test developers should continue to conduct traditional DIF studies.” This recommendation; however, does not address the issue that there is within group heterogeneity within such comparison groups and that DIF results do not apply to all examinees within the manifest groups. One option was illustrated in this study. That is, the use of LC membership as a DV and the analysis of DIF using traditional DIF detection methods once LCs have been identified in the test data structure. This may help address some of the issues raised by DeMars and Lau (2011) related to the ability of mixture IRT models to identify DIF.
This study makes two key contributions to measurement comparability research. First, findings from this study suggest that sample homogeneity should not be assumed. Instead, WGH should be empirically evaluated as one of the first steps in the analyses of comparability across groups. This is because, when the data is found to be heterogeneous, findings may apply only to a subgroup of examinees and not to the entire group. For example, findings from this study revealed that LC3 was stronger at reading for information as compared to the other LCs. This LC contained various proportions of examinees that were female, and that came from the different examined countries in varying proportions; thus, suggesting that neither gender nor country was homogeneous.

Second, if heterogeneity is found, it is important to investigate its sources. This investigation should be conducted by examining how variables (e.g., teacher and instruction related variables among others) influence examinee sub-groups’ (i.e., LCs’) item response patterns. This type of analysis seeks to examine for which sub-groups particular findings apply and in so doing aims to place inferential bounds on the findings (Zumbo, 2009). Such approaches may lead to more meaningful, useful and applicable findings to examinees that are targeted by the assessment (Ercikan, 2009).

One limitation of this study was that only one booklet from PIRLS 2006 was examined (i.e., the Reader). Although examining the Reader was advantageous because it provided large sample sizes required for manifest and latent class DIF analyses and it allowed for investigation of publicly released items, this booklet may not have been representative of other booklets in PIRLS 2006. For example, by using data from the Reader, issues of complex designs associated with international large-scale assessments (e.g., the use of matrix sampling) were not examined. Although, matrix sampling may be advantageous because it reduces the amount of time individual students are tested, it also significantly reduces the number of students taking each item; this may affect the reliability of the test and the comparability of the items. As the Reader was not matrix sampled, these issues were
not examined and thus future research should address the issue of matrix sampling and complex designs in the investigation of WGH. Second, whereas two manifest DIF detection methods were used to investigate manifest DIF, only one method was used to examine latent class DIF, which did not allow for analyzing consistency across latent class DIF approaches. This was because although there are several manifest DIF methods that can be used for DIF analyses, to my knowledge, the DMD model is the only model that allows for the examination of: (1) polytomous and dichotomous items typically found in international assessments and (2) item difficulty and item discrimination parameters.

Third, the selection of variables to differentiate across LCs may not have been closely connected to the particular factors influencing examinee response patterns. That is, although variables were selected based upon previous research which suggested that they had an influence on reading achievement, it is possible that the investigation of a greater number and greater diversity of variables may have led to explaining the distinction across LCs more clearly. Future studies should extend the number of variables examined. Further research should also explore to what extent LC DIF is related to cognitive requirements or types of reading (e.g., the use of a non-Roman alphabet, the use of ideograms, right-to-left vs. left-to-right direction of reading) assessed in multi-lingual reading assessments.

Fourth, the selected countries differed greatly in ability, the two Mandarin-speaking countries, had significantly higher achievement scores as compared to the two Arabic-speaking countries. This raises generalizability questions related to whether within group heterogeneity would also have been found if countries with more similar achievement scores would have been examined. In particular, As suggested by DeMars (2010), there are problems associated with the use of groups (countries) with large differences in proficiency which may result in false detection of DIF due to inaccurate matching. This is particularly the case in shorter measures because this may lead to less reliable test
scores. In the study conducted by DeMars (2010), these problems were revealed in relation to the use of manifest DIF methods; however, these may potentially also generalize to the use of latent class DIF methods. Future research should thus be conducted using data from other countries and languages. For example, for the purposes of cross-validation, languages using a non-Roman alphabet (e.g., Japanese, Korean) and those that are further removed from English (e.g., Turkish, Hungarian) may be selected as these languages and related countries have been found to have large proportions of DIF and item misfit (Oliveri & von Davier, 2011). It is important to note when conducting future research that although results obtained in this study indicated that the measure was unidimensional for the examined countries; this may not be the case when assessing other countries or student subgroups. For example, in the assessment of English Language Learners and students with disabilities there may be possible factors such as unnecessary linguistic complexity of items that may lead to multi-dimensionality. Hence, dimensionality should be examined as an important step in future research.
3. Effects of Within Group Heterogeneity on Methods Used to Detect Manifest DIF

3.1. Introduction

This study examined the effects of within group heterogeneity on detecting differential item functioning (DIF). Manifest DIF methods based on manifest grouping variables such as gender or ethnic background have the underlying assumption that DIF favouring a particular manifest group (e.g., girls) can be generalized to all test-takers of the same manifest group (e.g., all girls) in the population. These generalizations are based on the within group homogeneity (WGH) assumption and are valid only to the degree to which the homogeneity assumption is met.

To date, several studies have been conducted to investigate the WGH assumption using real data (Cohen & Bolt, 2005; De Ayala et al., 2002; Webb et al., 2008). Such studies have investigated the presence of latent classes (LCs) in the test data structure. Results from these studies reveal that examinees grouped using manifest characteristics (e.g., gender) have heterogeneous response patterns; that is, multiple LCs were found within manifest groups. Currently however, few studies have been conducted to examine to what degree within group heterogeneity has an effect on accuracy of detection of manifest DIF. Simulation studies lend themselves well for investigating the effects of within group heterogeneity on DIF detection by allowing the researcher to manipulate different heterogeneity conditions and examining their effect on DIF detection.

To address this paucity in the literature, this simulation study investigated the following research question:

What is the effect of within group heterogeneity on accuracy of DIF detection using manifest DIF methods?

Accuracy was examined by investigating false positive (FP) and correction detection (CD) rates of DIF detection.
The first section in this chapter discusses the WGH assumption in previous studies that investigated the effects of within group heterogeneity on accuracy of DIF detection using real and simulated data. The second section describes the simulation study and provides an example to illustrate within group heterogeneity. The method section describes the procedures used to generate the data, the examined independent variables (IVs), and the methods used to analyze the data. This chapter concludes by presenting and discussing results of analyses conducted to investigate the research question.

3.2. Literature Review

3.2.1. Investigation of the WGH Assumption in Real Data

This section summarizes findings from previous research conducted to investigate the WGH assumption in real data (Cohen & Bolt, 2005; Sawatzky et al., 2009; von Davier & Yamamoto, 2004). Collectively, findings from these studies suggest that the WGH assumption associated with manifest DIF methods is not tenable when analyzing real data because examinees have heterogeneous item responses. Cohen and Bolt (2005) conducted a study using data from a university mathematics placement test to investigate to what degree manifest groupings based on gender had homogeneous response patterns. These authors analyzed the data using a latent class DIF approach to investigate the degree to which manifest characteristics (i.e., gender) were related to the LCs being (dis)advantaged by the items. These analyses resulted in two LCs; each LC contained differing proportions of males and females. The findings from this study imply that both males and females within the examined population had heterogeneous response patterns and that there was a weak relationship between gender and the characteristics influencing examinees’ item responses which were more closely related to the content area being assessed by the test.
In another study, Sawatzky et al. (2009) used a latent variable mixture model to investigate response patterns of adolescents’ quality of life using a life satisfaction scale. Results revealed that improved model-data fit estimates were obtained when the sample was regarded as consisting of subpopulations or LCs rather than a homogeneous group. Findings suggest that adolescents had heterogeneous response patterns and that to accurately capture examinees’ item responses, LCs needed to be taken into account. Similarly, improved model-data fit was obtained in a study conducted by von Davier and Yamamoto (2004) when a three latent class model was used to analyze examinees’ item response patterns as compared to a one-class model also suggesting heterogeneity in examinees’ item responses. Study findings suggest that assuming homogeneity when the test data structure is heterogeneous leads to lower model-data fit estimates.

3.2.2. Simulation Studies Examining Within Group Heterogeneity

Two previous simulation studies have been conducted to examine the effects of within group heterogeneity on the accuracy of manifest DIF detection methods to detect DIF. First, De Ayala et al., (2002) conducted a study that simulated a 30-item fixed test length and used six different manifest DIF detection methods to investigate CD and FP DIF rates when there was heterogeneity simulated in the test data structure. The employed methods were: likelihood ratio (Thissen, Steinberg, & Wainer, 1993), logistic regression (LR; Swaminathan & Rogers, 1990), Lord’s Chi-Square (LCS; Lord, 1980), the Mantel-Haenszel (MH) Chi-Square (Holland & Thayer, 1988), and the Exact Signed Area (ESA) and H statistic approaches (Raju, 1988). Second, these issues were further investigated in a study conducted by Samuelsen (2005) who used the MH Chi-Square (Holland & Thayer, 1988) method with a fixed length test of 20 items to investigate within group heterogeneity. In these studies, within group heterogeneity was simulated by varying item parameters within comparison groups. That is, two subgroups were created
within each compared sample and DIF was introduced to only one of the two subgroups. DIF was introduced by increasing the difficulty parameter ($\Delta b$) of the focal group.

The study conducted by De Ayala et al., (2002) consisted of 50 replications and used three levels of DIF (0%, 10% and 20%) and two DIF effect sizes: $\Delta b$ 0.3 and $\Delta b$ 1.0. In the study conducted by Samuelsen (2005), 100 replications were performed. It used the Rasch model for data generation and examined three levels of DIF (10%, 30% and 50%) and three DIF effect sizes: $\Delta b$ 0.4, $\Delta b$ 0.8 and $\Delta b$ 1.2.

This research revealed that higher proportions of within group heterogeneity led to a slight decrease in FP rates (Samuelsen, 2005). On the other hand, the study conducted by De Ayala et al. (2002) found no particular patterns between degree of within group heterogeneity and FP rates. These studies found that greater proportions of DIF (e.g., 10% vs. 20% DIF) and higher DIF magnitudes led to higher FP rates. These studies also found that greater degrees of within group heterogeneity led to lower correct DIF detection rates. To illustrate, results from the study conducted by De Ayala et al. (2002) revealed that when the comparison groups were homogeneous, the LR, MH and IRT-based approaches correctly identified DIF items 61%, 56% and 49% of the time, respectively for the $\Delta b$ 0.3 effect size condition. For the $\Delta b$ 1.0 effect size condition, CD rates were 90% or higher. When the comparison groups were homogeneous, the study conducted by Samuelsen (2005) resulted in correct DIF detection rates of 90% or higher with effect sizes of $\Delta b$ 0.8 and $\Delta b$ 1.2 across 10%, 30% and 50% DIF conditions; these rates decreased to 50% or lower with effect sizes of $\Delta b$ 0.4. On the other hand, with 50% within group heterogeneity, CD rates decreased to below null condition levels for each of the six methods used in the study conducted by De Ayala et al. (2002). Similarly, in the study conducted by Samuelsen (2005), when the comparison group had higher degrees of within group heterogeneity (i.e., the comparison groups had 60% within group homogeneity or lower), correct DIF detection rates decreased to 10% or lower with 0.4 effect sizes, below 30% with $\Delta b$ 0.8 effect sizes.
and 70% or lower with $\Delta b$ 1.2 effect sizes. Findings from these studies suggest that correct DIF detection rates were also affected by the following four factors: small sample sizes (i.e., 500 examinees or fewer), high proportions of DIF (i.e., 20% or higher), DIF with small effect sizes (i.e., $\Delta b$ 0.4 or lower), and unequal ability distributions (i.e., differences in mean ability distributions of 1.0).

3. 3. The Current Study

The present study built upon previous studies conducted to investigate DIF when there is within group heterogeneity simulated in the test data structure in three ways. First, this study used a 2PL model to generate dichotomous items. In so doing, it built upon the study conducted by Samuelsen (2005) which used the Rasch model for data generation. As shown by previous studies, the Rasch model results in less accurate estimates of fit across multiple-choice (Divgi, 1986), and constructed-response (Fitzpatrick et al., 1996) item types typically used in ELSAs. Thus, this study followed recommendations made in previous studies to opt for the use of a more flexible, 2PL model (Divgi, 1986; Fitzpatrick et al., 1996). Second, this study used a methodical inclusion of Type I error control conditions (0% DIF) as compared to the Samuelsen (2005) study which did not include a 0% DIF baseline condition. Third, this study used a test length of 40 items, which as shown by previous research may be influenced by confounding factors such as the contamination of the matching criterion to a lesser extent than shorter tests (Donoghue & Allen, 1993); that is, 20 and 30 item test lengths used by the De Ayala et al., (2002) and Samuelsen (2005) studies, respectively.

In this study, diversity across examinee groups related to opportunity to learn was used as an example to illustrate differences in within group heterogeneity. Opportunity to learn was used because it exemplifies the type of differences that may be encountered when examining international large-scale assessment data. Differences associated with opportunity to learn were described in study one; to illustrate, as compared to examinees from Chinese Taipei and Hong
Kong, examinees from Qatar and Kuwait received different instruction and exposure to the curriculum. Specifically, whereas the curricula from Kuwait and Qatar emphasized learning to decode accurately and read fluently in Grades 3 to 5, the curricula from Hong Kong and Chinese Taipei emphasized understanding, organizing information, discussing and criticizing reading material as well as forming literate citizens to think independently (Kennedy et al., 2007). Similarly, differences in opportunity to learn within countries may also introduce variability within examinee groups. As will be described below, the latter scenario was used in this study to contextualize the simulation study as well as for data generation.

3.4. Method

3.4.1. Data Generation

In this study, data was generated with different degrees of within group heterogeneity, ranging from 0% to 80%. In the 0% within group heterogeneity (baseline) condition, the reference and focal groups were both homogeneous; that is, all examinees within groups had similar levels of opportunity to learn. Differences in opportunity to learn between groups existed; however, leading to DIF in some items. Keeping with the comparison groups used in study one, this may be viewed as the comparison of Chinese Taipei versus Hong Kong. As described in the PIRLS 2006 Encyclopedia, the reading curriculum from these two countries emphasized learning similar reading skills such as focusing on comprehension, organizing information as well as discussing and criticizing reading material to form literate citizens (Kennedy et al., 2007). As outlined in the 2006 PIRLS Encyclopedia, the curriculum from Hong Kong focused more on reading comprehension skills as compared to the curriculum from Chinese Taipei. Thus, examinees from Hong Kong may have been better able to respond to questions requiring high levels of reading comprehension skills as compared to examinees from Chinese Taipei. This condition is illustrated in Figure 3.1. The focal group (e.g., Chinese Taipei)
is represented by data set 2 (DS2) and the reference group (e.g., Hong Kong) by data set 1 (DS1). DIF was simulated by using higher $b$ parameters for a proportion of items (e.g., 12 out of 40 items) in the focal group as compared to the reference group.

**Figure 3.1. 0% Within Group Heterogeneity**

<table>
<thead>
<tr>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-12</td>
</tr>
<tr>
<td>13-40</td>
</tr>
</tbody>
</table>

**Groups**

- Reference
  - DS1
- Focal
  - DS2

[DIF-free]  [Lower $b$ parameters]  [Higher $b$ parameters]
The remaining (20% to 80%) within group heterogeneity conditions were simulated by creating heterogeneity within the focal group. Thus, examinees within the focal group represented test-takers with differing degrees of opportunity to learn the test content. For example, within the focal group (e.g., Chinese Taipei), there may be two examinee subgroups. One sub-group that may have been disadvantaged by test items requiring knowledge of higher order reading comprehension skills because they attended schools wherein teachers needed to focus on more basic skills (e.g., decoding) at the expense of following the mandated curriculum that emphasizes higher order reading comprehension skills. This type of heterogeneity may arise for various reasons including teaching in schools located in more impoverished areas or schools that have a high concentration of second language learners or have a higher concentration of immigrant students. A second sub-group of examinees may have attended schools teaching the mandated curriculum and have response patterns similar to examinees from the reference group in Figure 3.1. These two examinee sub-groups are illustrated in Figure 3.2. The proportion of examinees belonging to each of the two subgroups may vary, thus creating various degrees of within group heterogeneity. To exemplify, two within group heterogeneity conditions are illustrated in Figure 3.2: a 20% and an 80% within group heterogeneity condition. In the illustration depicting the 20% within group heterogeneity condition, 80% of examinees are from the focal group and 20% of examinees are from the reference group. In the illustration showing the 80% within group heterogeneity condition, 20% of examinees are from the focal group and 80% are from the reference group.
With this context in mind, this study investigated the effects of within group heterogeneity on the performance of two DIF detection methods (LR and MH). These methods were chosen for two reasons: 1) they are examples of commonly used DIF detection methods and 2) they could be automatized in SPSS. IRT based L-H would have been an alternative (as was used in study one); however, the software did not lend itself for simulation studies. The LR and MH methods were used to examine percentages of FP and CD DIF rates when there is within group heterogeneity in the data. FP DIF detection rates investigated items that were incorrectly detected as DIF when they were simulated to be DIF-free. CD rates examined rates of correctly identifying the items that were designed to have DIF.
3.4.2. Simulation Factors

The following factors were manipulated in this study: (1) degree of within group heterogeneity, (2) proportion of DIF and (3) effect size of DIF. Based on previous research, these three factors affect DIF detection.

1. **Degree of within group heterogeneity.** Five levels of this condition were manipulated: 0%, 20%, 40%, 60% and 80% to indicate that there may be differing proportions of heterogeneity within groups. The 0% within group heterogeneity level was used as a baseline condition to indicate homogeneity of the two manifest groups. The remaining (20% to 80%) levels represented an increasingly higher degree of within group heterogeneity. As shown in Figure 3.2, they were manipulated by varying the proportion of examinees from DS1 and DS2 making up the focal group. The 20% within group heterogeneity condition had a lower proportion (20%) of examinees from the reference group as compared to examinees from the focal group (80%). In the 80% within group heterogeneity condition, there were lower proportions (20%) of examinees from the focal group as compared to the reference group. Across all within group heterogeneity conditions, the reference group was homogeneous and was composed of examinees from DS1 only.

2. **Proportion of DIF.** As in previous studies (De Ayala et al., 2002, Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994, 1996), three proportions of DIF were simulated: 0%, 15% and 30%. Although current studies have found higher proportions of DIF (Allalouf et al., 1999; Ercikan et al., 2004; Sireci & Berberoglu, 2000); particularly, in relation to investigating comparability of international ELSAs administered to examinees of various cultural and linguistic backgrounds, higher percentages were not examined based upon findings from previous simulation studies, which suggest that higher proportions of DIF may lead to reductions in power to detect DIF and higher Type I error rates (Narayanan & Swaminathan, 1994, 1996). The
0% DIF was included as a baseline condition wherein the Type I error rate was expected to be at or below the nominal level.

3. **DIF effect sizes.** As in previous simulation studies conducted to investigate DIF (Clauser et al., 1993; De Ayala et al., 2002; Penfield, 2001; Samuelsen, 2005), effect sizes in this study were manipulated based on the difference between item difficulty parameters (Δb) for the comparison groups. Specifically, the item was made more difficult for the focal group by adding Δb to that group. Please recall that an IRT model was used for data generation, thus a Δb refers to a change in IRT “b” parameters. Three DIF effect sizes (or levels of simulated DIF) were generated: Δb 0.35 (Camilli & Shepard, 1987), referred to as level 1 DIF, Δb 0.48 (Swaminathan & Rogers, 1990), referred to as level 2 DIF, and Δb 0.64 (Swaminathan & Rogers, 1990), referred to as level 3 DIF. These effect sizes were selected based upon parameters used in the study conducted by Swaminathan and Rogers (1990), wherein a Δb of 0.48 and Δb 0.64 represented moderate and large effect sizes, respectively. Moreover, a Δb of 0.35 was used based on the study conducted by Camilli & Shepard (1987), wherein DIF of this magnitude was considered to have a small effect size.

Table 3.1 summarizes the factors and levels that were manipulated in this study, including degree of within group heterogeneity, proportion and effect size of DIF. This led to a 5 x 2 x 3 crossed factorial design; in addition there was 1 no-DIF condition that represented the baseline condition (0% within group heterogeneity). Thus, there were a total of 31 cells, which were analyzed using each of the following two manifest DIF detection methods: LR and MH.
Table 3.1

Overview of Manipulated IVs

<table>
<thead>
<tr>
<th>DIF: % &amp; Magnitude</th>
<th>0%</th>
<th>20%</th>
<th>40%</th>
<th>60%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δb 0.35 (Level 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δb 0.48 (Level 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δb 0.64 (Level 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δb 0.35 (Level 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δb 0.48 (Level 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δb 0.64 (Level 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

100 replications of each of the cells shown in Table 3.1 were conducted; this number of replications is consistent with previous simulation studies that investigated the performance of DIF under various conditions (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994, 1996; Samuelsen, 2005).

3.4.3. Simulation Procedures

Data for this study was generated for a fixed test length of 40 binary items. This test length was selected because it may lead to more stable parameter estimations and lesser influence from confounding factors such as the contamination of the matching criterion than shorter tests. Moreover, as noted in previous simulation studies conducted by Jodoin and Gierl (2001), Narayanan and Swaminathan (1994, 1996), and Rogers and Swaminathan (1993), which
also used a 40-item test length, this number of items represents a short but reliable standardized test of achievement.

Data was generated using WinGen2 (Han, 2007) because this software has been designed to run on Windows 32 or 64 bit operating systems, which enables faster and more effective simulation of item response data for large numbers of examinee sample sizes (as large as 100,000,000) than alternative programs which typically generate data for only 4,000 examinees (Han & Hambleton, 2007). It also enables saving of parameter estimates generated at three stages of data simulation: (1) generating ability parameter values, (2) generating item parameter values, and (3) simulating item response data (Han & Hambleton, 2007) which allowed for more detailed specification of the parameters that were simulated.

Two data sets (DS1 and DS2) consisting of 200,000 cases each was generated. This population size was selected to allow for conducting 100 replications consisting of a maximum of 1,000 examinees per level of each IV examined. Thus, this population size enabled the random sampling of examinees to be used for data analysis to come from a population that had the same characteristics. That is, to allow for within factor analyses of the simulated data to be conducted using subsets of data randomly selected from the generated population. One data set, DS1, represented the reference group as well as a sub-group of examinees in the focal group. Another data set, DS2, represented the focal group when there was no heterogeneity in the test data and a sub-group of focal group examinees when there was heterogeneity in the test data structure. To create differing degrees of heterogeneity in the focal group, examinees from DS1 and DS2 were merged according to the proportion of heterogeneity condition being examined (e.g., 20%, 40%, 60%, or 80%). In the 0% condition (the baseline condition) the sample was homogeneous; there was no within group heterogeneity and thus no examinees from the focal group.
The following sample sizes were used: the reference group had 1,000 examinees and the focal group also had 1,000 examinees. Sample sizes of 1,000 examinees were selected based upon previous studies with real (Hambleton & Rogers, 1989) and simulated data (Clauser et al., 1993; Narayanan & Swaminathan, 1994; Penfield, 2001), which revealed that larger sample sizes (i.e., 1,000 examinees per group) should be used to not confound identification of DIF with issues of sample size. Ability distributions for the two groups (DS1 and DS2) were designed to have a normal distribution with mean of zero and standard deviation of one.

Two factors in this simulation study were designed to correspond to study one in this dissertation. First, this study used a 2PL model to generate data, which was the model used to scale PIRLS 2006 dichotomous test items. Second, IRT item parameters for the reference group were taken from study one and were used for the first 21 items in this study. The remaining 19 parameters were randomly generated.

DIF items were selected upon findings from previous studies (Hambleton & Rogers, 1989; Narayanan & Swaminathan, 1994) which suggest that item type had a significant effect on DIF detection rates. Two criteria were used for item selection. First, items with low discrimination \(a<1.0\) were excluded because as suggested in the study conducted by Hambleton and Rogers (1989), items with low discrimination often lead to unstable DIF statistics. Second, items with moderate difficulty parameters \(|b|>0.6\) were selected based upon findings from the study conducted by Narayanan and Swaminathan (1994), wherein items with difficulty parameters below -0.6 and above 0.6 were more difficult to identify as DIF.

Table 3.2 lists item parameters used in this study. The first three columns list the discrimination \(a\) and difficulty \(b\) parameters used from study one and the last three columns present the randomly generated discrimination \(a\) and difficulty \(b\) parameters. DIF was simulated for 15% and 30% of items depending on the condition being manipulated; items used
in the 15% condition are marked by superscript (\(^a\)). In addition to those items, items used in the 30% condition are marked by superscript (\(^b\)).

**Table 3.2**

*Item Parameters*

<table>
<thead>
<tr>
<th>IRT item parameters from study 1</th>
<th>Randomly selected item parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item #</td>
<td>(a) parameter</td>
</tr>
<tr>
<td>--------</td>
<td>------------------</td>
</tr>
<tr>
<td>1</td>
<td>0.477</td>
</tr>
<tr>
<td>2</td>
<td>1.093</td>
</tr>
<tr>
<td>3</td>
<td>0.740</td>
</tr>
<tr>
<td>4</td>
<td>0.675</td>
</tr>
<tr>
<td>5</td>
<td>1.550</td>
</tr>
<tr>
<td>6</td>
<td>0.897</td>
</tr>
<tr>
<td>7</td>
<td>1.013</td>
</tr>
<tr>
<td>8</td>
<td>0.740</td>
</tr>
<tr>
<td>9</td>
<td>1.521</td>
</tr>
<tr>
<td>10</td>
<td>0.256</td>
</tr>
<tr>
<td>11</td>
<td>0.678</td>
</tr>
<tr>
<td>12</td>
<td>0.568</td>
</tr>
<tr>
<td>13</td>
<td>0.724</td>
</tr>
<tr>
<td>14</td>
<td>0.593</td>
</tr>
<tr>
<td>15</td>
<td>0.806</td>
</tr>
<tr>
<td>16</td>
<td>2.037</td>
</tr>
<tr>
<td>17</td>
<td>1.732</td>
</tr>
<tr>
<td>18</td>
<td>1.105</td>
</tr>
<tr>
<td>19</td>
<td>1.271</td>
</tr>
<tr>
<td>20</td>
<td>0.998</td>
</tr>
<tr>
<td>21</td>
<td>0.382</td>
</tr>
</tbody>
</table>

*Notes:* IRT=Item Response Theory  
\(^a\)= items used in 15% DIF condition  
\(^b\)= items used in 30% DIF condition
3.4.4. Analysis of Simulated Data

Resulting samples were manipulated under the factors summarized in Table 3.1. These samples were analyzed using two DIF detection methods: LR and MH. Each of these methods is described next.

3.4.5. Analysis of Manifest DIF

FP and CD rates were examined separately for each manifest DIF method analyzed: LR and MH. Critical $\chi^2$ values were set at 5% and 1% significance levels.

**LR DIF.** The LR method (Swaminathan & Rogers, 1990) is based on a logistic regression equation, and involves regressing the total score, group membership (i.e., language, gender, ethnicity, or race), and the interaction between total score and group membership. The total score is the sum of examinees’ responses to test items. The regression equation is conducted in a step-wise fashion resulting in a $\chi^2$ based test of significance and an effect size (Nagelkerke $R^2$) obtained for each step. Differences between $R^2$ values with the addition of each new step in the model provide a measure of variance accounted for by the added variable (Zumbo, 1999). SPSS was used to examine DIF using the LR method. Critical $\chi^2$ values were set at 5% and 1% significance levels.

**MH DIF.** The second manifest DIF method used was MH. This method uses a 2 x 2 contingency table to calculate total test score categories using counts of examinees from two groups (i.e., the focal and reference groups) who are matched on the total score. Such groups are divided into levels dependent upon their likelihood of answering an item correctly; DIF is identified when different examinee groups have differing likelihoods of scoring correctly on an item (Clauser & Mazor, 1998). These analyses result in a $\chi^2$ distributed statistic with one degree of freedom as well as a log odds ratio (LOR) or difference measure ($\Delta MH - \alpha$). SPSS was used to examine DIF using the MH method. Critical $\chi^2$ values were set at 5% and 1% significance levels.
The total score was used for matching examinees on ability. In this type of matching (referred to as thin matching), the examined set of items is pooled across total score levels. Thin matching has been found to perform better than thick matching with 40-item test lengths (Donoghue & Allen, 1993).

3.5. Results

Results of DIF analysis using manifest detection methods with respect to FP and CD DIF rates are summarized in this section in three parts. First, results of the baseline (0% DIF) condition are presented. Next, results of mean percentage FP DIF rates followed by results of CD DIF rates are summarized. Results of FP rates are presented first in order to establish nominal or below nominal FP rates to help make appropriate interpretations of CD rates (French & Maller, 2007).

3.5.1. The 0% DIF Condition

The 0% DIF condition serves as a control condition wherein Type I error rates should be at or below nominal levels. At \( p \)-value=.01, the 0% DIF condition resulted in 1% FP DIF rates for the MH and LR methods. At \( p \)-value=.05, the 0% DIF condition resulted in 4% and 5% FP DIF rates for the MH and LR methods, respectively. These results indicate that Type I error control conditions were within nominal rates.

3.5.2. False Positive DIF Detection Rates

FP DIF detection rates indicate the percentage of items that were simulated to be DIF-free and were incorrectly detected as functioning differentially. For example, in the 30% DIF condition, 12 out of 40 items were designed to have DIF and 28 items were designed to be DIF-free. For each cell in Table 3.1, this percentage was calculated per iteration over the 100 iterations. A mean percentage was estimated across the 100 iterations to obtain the FP rate per cell. The mean percentage FP rates indicate the proportion of items (out of the 34 DIF-free
items) that were incorrectly detected as DIF. FP DIF detection rates for the MH and LR methods are summarized in Tables 3.3 and 3.4, respectively. In these tables, the first column indicates the percentage of manipulated DIF (i.e., 15% or 30% DIF). The second column indicates the studied degree of within group heterogeneity (labelled WGH). Columns labelled .05 and .01 display FP rates corresponding to critical $\chi^2$ test values of .05 and .01, respectively. Numbers in bold indicate FP DIF rates that are above the 5% nominal rate.

Table 3.3

| Results of Mean Percentage False Positive Rates Using the MH Method of DIF Detection |
|---------------------------------|--------|--------|--------|--------|--------|--------|
| DIF    | WGH   | Level 1 | Level 2 | Level 3 | Level 1 | Level 2 | Level 3 |
|        | WGH   | .05     | .01     | .05     | .01     | .05     | .01     |
| 15%    | 0%    | 8       | 2       | 12      | 4       | 20      | 8       |
|        | 20%   | 7       | 2       | 9       | 3       | 13      | 4       |
|        | 40%   | 6       | 1       | 6       | 2       | 9       | 3       |
|        | 60%   | 4       | 1       | 5       | 1       | 6       | 2       |
|        | 80%   | 4       | 1       | 4       | 1       | 4       | 1       |
| 30%    | 0%    | 18      | 7       | 28      | 13      | 45      | 26      |
|        | 20%   | 12      | 4       | 20      | 8       | 32      | 16      |
|        | 40%   | 9       | 3       | 13      | 4       | 21      | 8       |
|        | 60%   | 5       | 1       | 7       | 2       | 11      | 4       |
|        | 80%   | 4       | 1       | 5       | 1       | 5       | 2       |

Notes: WGH = degree of within group heterogeneity
.05 = statistically significant DIF results at $p$-value=.05
.01 = statistically significant DIF results at $p$-value=.01
Table 3.4

*Results of Mean Percentage False Positive Rates Using the LR Method of DIF Detection*

<table>
<thead>
<tr>
<th>Simulated Level of DIF</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level 1</td>
<td>Level 2</td>
<td>Level 3</td>
</tr>
<tr>
<td>DIF (WGH&lt;sub&gt;e&lt;/sub&gt;)</td>
<td>.05</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td>0%</td>
<td>8</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>20%</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>40%</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>60%</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>80%</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0%</td>
<td>16</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>20%</td>
<td>11</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>40%</td>
<td>9</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>60%</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>80%</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

*Notes:* WGH<sub>e</sub> = degree of within group heterogeneity
.05 = statistically significant DIF results at p-value=.05
.01 = statistically significant DIF results at p-value=.01

As shown in Tables 3.3 and 3.4, the 0% within group heterogeneity condition yielded above nominal rates across 15% and 30% DIF conditions at p-value=.05. Similar FP rates (15% and 22% FP rates for MH and LR) were found in a study conducted by Navas-Ara and Gómez-Benito (2002), wherein similar percentages of DIF (40% DIF) and DIF magnitude (Δb=.075) were manipulated. The use of a stricter p-value (.01) resulted in lower FP rates for the LR and MH DIF detection methods. To illustrate, in the homogeneous condition, when using p-value=.01 as compared to p-value=.05, the number of conditions with above nominal rates
decreased to four out of six (67%) for the MH method and three out of six (50%) for the LR method. Despite this decrease, the number of conditions with above nominal FP rates remained high.

Analysis of the conditions wherein group heterogeneity was manipulated indicates that higher degrees of heterogeneity led to a decrease in FP DIF rates. For example, as shown in Table 3.3, the 15% Level 3 DIF condition at $p$-value=.05 resulted in 20% FP rates when there was 0% heterogeneity; these rates decreased to 4% in the 80% heterogeneity condition. This pattern was observed for the MH method as well as the LR method. These results are consistent with the studies conducted by Samuelsen (2005) wherein there was a decrease in FP rates associated with increased rates of within group heterogeneity.

Results also indicated higher FP rates associated with a higher degree of DIF across all within group heterogeneity conditions. These results are consistent with previous studies examining the homogeneous condition (French & Maller, 2007; Hidalgo-Montesinos & Gómez-Benito, 2003) as well as those investigating within group heterogeneity (Samuelsen, 2005).

3.5.3. Correct DIF Detection Rates

Correct DIF detection rates are mean percentage rates of correct identification of known DIF items. Mean percentage for CD rates was calculated in the same way as for FP rates. That is, a percentage of correctly identified items was calculated per iteration across the 100 iterations. Mean percentage CD rates was estimated by averaging across the 100 iterations. CD DIF results for the MH and LR methods are summarized in Tables 3.5 and 3.6, respectively. In these tables, the first column indicates the percentage of manipulated DIF (i.e., 15% or 30% DIF). The second column indicates the studied degree of within group heterogeneity (labelled WGHe). Columns labelled .05 and .01 display CD rates corresponding to critical $\chi^2$ test values of .05 and .01, respectively.
### Table 3.5

**Results of Mean Percentage Correct Detection Rates Using the MH Method of DIF Detection**

<table>
<thead>
<tr>
<th>DIF</th>
<th>WGHe</th>
<th>Simulated Level of DIF</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Level 1</td>
<td>Level 2</td>
<td>Level 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>15%</td>
<td>0%</td>
<td>86</td>
<td>99</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20%</td>
<td>69</td>
<td>92</td>
<td>99</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40%</td>
<td>42</td>
<td>73</td>
<td>93</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60%</td>
<td>23</td>
<td>40</td>
<td>66</td>
<td>42</td>
</tr>
<tr>
<td>0%</td>
<td>80%</td>
<td>8</td>
<td>8</td>
<td>14</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20%</td>
<td>64</td>
<td>85</td>
<td>98</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40%</td>
<td>43</td>
<td>68</td>
<td>91</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60%</td>
<td>25</td>
<td>47</td>
<td>71</td>
<td>49</td>
</tr>
<tr>
<td>0%</td>
<td>80%</td>
<td>1</td>
<td>5</td>
<td>23</td>
<td>42</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20%</td>
<td>14</td>
<td>23</td>
<td>42</td>
<td>21</td>
</tr>
<tr>
<td>0%</td>
<td>80%</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>15</td>
<td>4</td>
</tr>
</tbody>
</table>

**Notes:**
- WGH = degree of within group heterogeneity
- .05 = statistically significant DIF results at *p*-value=.05
- .01 = statistically significant DIF results at *p*-value=.01
Table 3.6

Results of Mean Percentage Correct Detection Rates Using the LR Method of DIF Detection

<table>
<thead>
<tr>
<th>Simulated Level of DIF</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DIF</strong></td>
<td>WGH: .05</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td><strong>15%</strong></td>
<td>0%</td>
<td>77</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>60</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>34</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>60%</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td><strong>30%</strong></td>
<td>0%</td>
<td>54</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>60%</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

Notes: WGH = degree of within group heterogeneity
.05 = statistically significant DIF results at $p$-value=.05
.01 = statistically significant DIF results at $p$-value=.01

As shown in Tables 3.5 and 3.6, consistent with previous research (De Ayala et al., 2002), CD rates increased with higher levels of simulated DIF magnitude. To illustrate, higher percentages of CD rates (100%) were found for the MH and LR methods in the 15% Level 3 DIF condition at $p$-value=.05. These rates decreased with lower levels of simulated DIF. For example, in the level 1 DIF condition, the MH and LR methods resulted in 86% and 77% CD rates, respectively.

In addition, the following three factors led to a decrease in CD rates: (1) the use of a stricter statistical test (i.e., $p$-value=.05 vs. $p$-value=.01), (2) greater percentages of simulated
DIF (i.e., 15% DIF vs. 30% DIF), and (3) greater proportions of within group heterogeneity. These three factors are described in more detail next.

First, power decreased when going from $p$-value=.05 to $p$-value=.01 across all examined conditions and across all group heterogeneity levels (see Tables 3.5 and 3.6). These results are expected and are consistent with a study conducted by van der Flier, Mellenbergh, Adèr and Wijn (1984) wherein CD rates decreased in relation to the use of more stringent statistical significance test values ($p$-value=.05 vs. $p$-value=.01).

Second, across all conditions and group heterogeneity levels, lower CD rates were associated with greater percentages of DIF (i.e., 30% DIF as compared to 15% DIF). These results were not surprising and correspond to findings from previous studies (French & Maller, 2007; Hidalgo-Montesinos & Gómez-Benito, 2003). For the 0% WGH condition, these findings may be related to an increased contamination of the matching criterion due to greater proportions of DIF (Zenisky, Hambleton, & Robin, 2003). However, for the 20% to 80% within group heterogeneity conditions, the relationship between contamination of the matching criterion and proportion of heterogeneity is less clear. To illustrate, as shown in the study conducted by Samuelsen (2005) wherein contamination of the matching criterion was evaluated across different levels of heterogeneity, the use of purification led to minimal differences in correct DIF detection rates pre- and post-purification. Further research is needed in this area and falls beyond the scope of this study.

Three, for the two examined DIF detection methods, an inverse relationship between correct DIF detection rates and heterogeneity was observed. Specifically, higher levels of sample heterogeneity led to lower CD rates. For example, as shown in Table 3.5, the 30% level 3 DIF condition (at $p$-value=.01) resulted in 93% CD rates when there was 0% WGH and decreased to 4% in the 80% within group heterogeneity condition.
Findings that greater degrees of within group heterogeneity lead to decreased CD rates are consistent with previous studies. To illustrate, results from the study conducted by De Ayala et al. (2002) revealed that when the comparison groups were homogeneous, the LR, MH and IRT-based approaches correctly identified DIF items 61%, 56% and 49% of the time, respectively for the $\Delta b$ 0.3 effect size condition. For the $\Delta b$ 1.0 effect size condition, CD rates were 90% and higher. With 50% heterogeneity, CD rates decreased to below null condition levels for each of the six methods investigated. Similarly, the study conducted by Samuelsen (2005) resulted in correct DIF detection rates of 90% or higher with effect sizes of $\Delta b$ 0.8 and $\Delta b$ 1.2 across 10%, 30% and 50% DIF conditions; whereas with higher degrees of sample heterogeneity (i.e., 60% heterogeneity), correct DIF detection rates decreased to 10% or lower with $\Delta b$ 0.4 effect sizes, below 30% with $\Delta b$ 0.8 effect sizes and 70% or lower with $\Delta b$ 1.2 effect sizes.

3.6. Summary and Discussion

In this simulation study, the effects of various proportions of within group heterogeneity on FP and CD rates associated with the LR and MH manifest DIF detection methods were investigated. DIF was investigated using two $\chi^2$ significance levels (.05 and .01), three levels of DIF and two proportions of DIF (15% and 30%). Within group heterogeneity rates ranged from 0% (the baseline condition) to 80%.

Manipulation of the three factors simulated in this study led to the following results. First, as expected, the use of higher levels of DIF led to higher CD rates. Second, the use of higher proportions of DIF led to lower CD rates. Third, higher rates of within group heterogeneity led to a decrease in CD rates. Reductions in CD rates occurred to the extent that with greater proportions of within group heterogeneity (e.g., at or above 60% heterogeneity), CD rates were below nominal.
Finding that higher proportions of DIF led to lower CD rates suggests that analyses of the effects of the contamination of the matching criterion should be conducted. Contamination of the matching criterion occurs when the total test score that is used to match examinees on ability includes differentially functioning items. Results from previous studies reveal that the effects of contamination are higher when tests have greater percentages of DIF (Zenisky et al., 2003). To address this issue, purification of the matching criterion, which involves analyzing DIF using an iterative procedure wherein items identified as DIF in the first stage are removed from the matching criterion at subsequent stage(s) has been proposed. Findings from previous studies suggest that purification leads to reducing Type I error rates and enhancing CD rates (Navas-Ara & Gómez-Benito, 2002; Zenisky et al., 2003). In contrast, previous studies have also shown limitations associated with the purification procedure. For example, a simulation study conducted by French and Maller (2007) revealed that purification led to increased Type I error rates. It is also unknown to what degree the removal of items when purifying the matching criterion would affect the reliability of the matching criterion and the subsequent DIF results. Further, the use of purification is not a well established procedure in practice; particularly, with either the LR method (French & Maller, 2007) or the examination of within group heterogeneity. Given mixed results associated with purification, this procedure was not used in this study. However, to clarify the role of purification effects on accuracy of DIF detection when there is within group heterogeneity simulated in the test data structure, future studies should be conducted.

Finding that greater degrees of within group heterogeneity lead to reduced correct DIF detection rates is consistent with previous studies (De Ayala et al., 2002; Samulsen, 2005). This finding implies that when there are greater degrees of within group heterogeneity, claims regarding the comparability of tests may lose their meaning and accuracy because the test may have higher proportions of DIF than those actually detected by manifest DIF analyses. This finding may also
explain why the use of manifest DIF detection methods in study one revealed a lower number of detected differentially functioning items as compared to the use of the latent class DIF approach.

In this study, within group heterogeneity was narrowly defined. It was simulated using two scenarios. One (the baseline condition) was a homogeneous condition wherein the reference and focal groups were both homogeneous. In the second type, within group heterogeneity was simulated for the focal group by simulating an examinee sub-group that had response patterns akin to the reference group. These two scenarios, however, do not encompass the different kinds of within group heterogeneity that may arise when analyzing large-scale assessment data. Study one is discussed here to exemplify additional scenarios where within group heterogeneity may arise. To illustrate, within group heterogeneity may arise due to differences in reading comprehension skills, as was shown in study one, one LC may have more experience with reading fiction compared to non-fiction texts, or may have more skills to respond to questions requiring higher-order reading skills. Alternatively, within group heterogeneity may occur within the context of assessing English Language Learners (ELL) and students with learning disabilities. For example, within group heterogeneity in ELLs may arise as these students come from different cultures; have different levels of English proficiency and different background characteristics. These two scenarios as well as those described in the study conducted by DeMars and Lau (2011) related to simulating multidimensionality and fewer DIF items should serve as context for future research studies on the simulation of within group heterogeneity.

Further research should also be conducted to investigate the effects of within group heterogeneity under a greater number of experimental conditions. In this study, three factors were manipulated: percentage of DIF, degree of DIF and degree of within group homogeneity. To increase generalizability of results other factors should be manipulated. These factors include varying sample size, test length, proportions of DIF, and level and types of within group heterogeneity influencing DIF detection.
4. Conclusion

The purpose of this dissertation was to investigate the assumption of within group homogeneity (WGH) and the effects of the violation of this assumption on DIF detection. To this end, two studies were conducted. The first study investigated this assumption using data from the PIRLS 2006 Reader administered to examinees from Chinese Taipei, Hong Kong, Kuwait and Qatar. The second study examined the effects of the violation of within group heterogeneity on correct detection (CD) and false positive (FP) DIF rates associated with two manifest DIF detection methods: Logistic Regression (LR) and Mantel-Haenszel (MH). These two studies were connected as they both investigated the effects of within group heterogeneity associated with manifest DIF detection methods within international large-scale assessment contexts. Jointly, these studies contributed to addressing the lack of research associated with examining the WGH assumption when using manifest DIF detection methods.

This chapter highlights key findings and contributions each of the two studies makes to the literature in educational measurement in general and on measurement comparability research in particular. It also discusses implications and limitations of each of the two studies. Moreover, suggestions for future research are given.

4.1. Contributions

This dissertation makes three major contributions to the measurement literature in general and to the investigation of DIF and its sources using international large-scale assessment data in particular. First, it contests the assumption of WGH using an international assessment context and presents evidence to demonstrate how the WGH assumption may jeopardize the validity of DIF investigations. This is an important contribution given that the majority of DIF research in educational large-scale assessments (ELSAs) has been conducted using manifest DIF methods under the WGH assumption. This has been the case despite the diversity of examinees participating in
international assessments. In fact, to my knowledge, this is the only instance of a study conducted to examine WGH using international large-scale assessment data.

A second contribution is related to the examination of sources of DIF using cognitive and teacher background variables, in addition to typically used manifest characteristics (e.g., gender, language of test, country). Findings from study one suggest that manifest characteristics typically used to group examinees (e.g., gender or country) may not be the most significant variables explaining or capturing variation in examinees’ item responses. Instead, findings from study one suggest that other factors, such as examinee ability groupings, should be investigated in examining measurement comparability. These types of analyses acknowledges that there are key factors such as psychosocial and cultural variables (Zumbo & Gelin, 2005) as well as community and background variables (Muthén, 1989, Muthén, Kao, & Burstein, 1991) that may impact the testing situation and that should be included in measurement comparability investigations.

Third, the two studies conducted in this dissertation served to elucidate the effects of assuming within group heterogeneity on the detection of DIF. Findings from study one revealed that different conclusions (i.e., regarding the number of differentially functioning items) were obtained depending on whether DIF was investigated for manifest groups or whether it was examined for latent groups. To illustrate, in study one, the use of a manifest DIF approach led to the detection of a lower number of items with lower effect sizes as compared to the use of a latent class DIF approach. As the actual proportion of DIF and related effect size was not known in study one, study two was conducted to further examine these issues in a simulation. Study two findings revealed that the use of manifest DIF detection methods with data that has a heterogeneous structure leads to lower CD rates. Findings from study two suggest that when using real data, the WGH assumption needs to be evaluated prior to making claims regarding the comparability of a measure because if the sample is heterogeneous, DIF may be under detected.
4.2. Implications

Findings from this dissertation have three important implications. The first implication is the generalization of findings that are applicable to sub-groups to an entire group. The second is related to the effects of within group heterogeneity in relation to decreased CD rates. Last, findings from this research have an implication on how the practices of test validation and measurement comparability analyses are conducted.

One implication of study findings and of the assumption of WGH is related to generalizing findings that apply only to sub-groups to the entire group. As mentioned in the introduction, when data analyses are conducted using manifest variables (e.g., gender), there is an underlying assumption that the examinee group (e.g., girls) has homogeneous response patterns. This assumption leads to generalizing findings to the entire population (e.g., all girls), and may lead to stereotypes and inaccurate claims regarding the sample. Thus, as stated by Ercikan (2009) and Zumbo (2009) inferences made from a population may be valid for one sub-group (high-performing females) but not for another (low-performing females); or for one setting (e.g., females living in a large city in Canada) and not for another (females living in Canadian rural communities). The assumption of WGH may lead to results that are inaccurate and are not applicable to the examinees targeted by the assessment; this was illustrated in study one as findings that applied to one LC did not apply to other LCs.

As generalizability is one of the six aspects of the validation process described by Messick (1989; 1995), empirical investigations of sample heterogeneity should be conducted to increase the validity of inferences made from test results. As stated by Ercikan (2009), whether research provides useful direction for practice and policy depends on the degree to which findings apply to key sub-groups that are homogeneous.

Second, findings from the two studies conducted in this dissertation elucidated the effects of within group heterogeneity in relation to correct manifest DIF detection rates. Specifically, findings
from study two revealed that higher rates of heterogeneity led to lower CD rates. As highlighted by Samuelsen (2008), sample heterogeneity may lead to obscuring the true degree of DIF when conducting manifest DIF analyses. As greater degrees of heterogeneity lead to having lower degrees of DIF detected in the measure, this may lead to developing biased tests as well as threatening the comparability of the measure for examinee sub-groups.

The third implication is related to the assumptions made when conducting test comparability analyses. Specifically, findings from study one revealed that the analyzed data was heterogeneous, had typical procedures used to detect manifest DIF in international assessments been used, this would not have been detected. Findings from study two revealed that the use of manifest DIF methods with data that is heterogeneous leads to reduction of correct DIF detection rates. These findings suggest that the current practices and assumptions being employed in investigating measurement comparability when conducting bias analyses need to be revised to incorporate the empirical analysis of the WGH assumption.

4.3. Limitations

Jointly, these studies have some limitations which are related to the generalizability of findings from the studies conducted. Specifically, the first study used data from only one booklet. As mentioned, this booklet may not have been representative of other booklets in PIRLS 2006. Thus, findings stemming from the use of this booklet may not generalize to all of PIRLS 2006. Moreover, the first study used data from one international large-scale assessment: PIRLS. This assessment focused on reading literacy. Findings may have been different if other curricular areas had also been examined (e.g., mathematics or science). For example, findings from previous studies suggest that reading has a higher proportion of DIF as compared to science and mathematics (Oliveri & von Davier, 2011), and suggest that reading literacy depends on the quality of translations more than
international comparisons using translated mathematics and science items (Grisay et al., 2009). Thus, future studies should be conducted to examine the effects of heterogeneity on other curricular areas.

In the second study, heterogeneity was created by simulating a focal group composed of two subgroups with different opportunities to learn. However, there may be different contexts for within group heterogeneity to arise such as differences in socio-economic status or language background among other factors. This simulation design did not encompass the different types of within group heterogeneity that may be encountered when analyzing large-scale assessment data. Thus, future research should be conducted to broaden the definition of within group heterogeneity.

4.4. Future Research

As this is an emerging area of research, there are several areas that need further investigation. Two research areas are described in this section. One is related to further research using real data. The second is associated with further studies conducted using simulated data.

In relation to real data studies, one area for future research is related to the use of background questionnaires to further understand sources of heterogeneity in examinee item responses. To illustrate, the PIRLS 2006 teacher questionnaire was used in study one to examine teacher background variables associated with differences in instructional approaches. This evaluation was conducted to further investigate variations in examinee sub-groups’ item responses. Further research should be conducted using alternative measures (e.g., PISA and TIMSS), background questionnaires (e.g., administered to teachers, parents and students), and variables (e.g., motivation, socio-economic status). Moreover, other studies should be conducted to investigate the generalizability of findings from study one with examinees from other countries and cultural backgrounds.

Further investigation of within group heterogeneity using simulated data should be conducted using other manifest DIF detection methods (e.g., to investigate robustness of IRT in
relation to the manipulation of different levels of sample heterogeneity). Moreover, the simulation study conducted in this dissertation used an equal sample size for the focal and reference groups, other simulation studies should be conducted by varying sizes of the reference and focal groups, as well as varying test lengths (short, medium and long tests) and varying the proportion of DIF items. Further, in study two, the unit of analysis was the item-level, other simulation studies may focus on examining the effects of varying within group heterogeneity on the total score. A change in the unit of analysis to the total score level may allow investigation of the effects of sample heterogeneity on consequences made from tests which are typically based on total test scores.
REFERENCES


Cook, L. (July, 2006). *Practical considerations in linking scores on adapted tests*. Keynote address at the 5th international meeting of the International Test Commission, Brussels, Belgium.


Zumbo, B. D. (2007a). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly, 4*, 223-233.


An Unbelievable Night

by Franz Hohler

Anina was ten years old, so even half asleep she could find her way from her room to the bathroom. The door to her room was usually open a crack, and the nightlight in the hallway made it light enough to get to the bathroom past the telephone stand.

One night, as she passed the telephone stand on her way to the bathroom, Anina heard something that sounded like a quiet hissing. But, because she was half asleep, she didn’t really pay any attention to it. Anyway, it came from pretty far away. Not until she was on her way back to her room did she see where it came from. Under the telephone stand there was a large pile of old newspapers and magazines, and this pile now began to move. That was where the noise was coming from. All of a sudden the pile started to fall over – right, left, forwards, backwards – then there were newspapers and magazines all over the floor.
Anina could not believe her eyes as she watched a grunting and snorting crocodile come out from under the telephone stand.

Anina was frozen to the spot. Her eyes wide as saucers, she watched the crocodile crawl completely out of the newspapers and slowly look around the apartment. It seemed to have just come out of the water because its whole body was dripping wet. Wherever the crocodile stepped, the carpet under it became drenched.
The crocodile moved its head back and forth letting out a loud hissing sound. Anina swallowed hard, looking at the crocodile’s snout with its terribly long row of teeth. It swung its tail slowly back and forth. Anina had read about that in “Animal Magazine”—how the crocodile whips the water with its tail to chase away or attack its enemies.

Her gaze fell on the last issue of “Animal Magazine,” which had fallen from the pile and was lying at her feet. She got another shock. The cover of the magazine used to have a picture of a big crocodile on a river bank. The river bank was now empty!

Anina bent down and picked up the magazine. At that moment the crocodile whipped his tail so hard that he cracked the big vase of sunflowers on the floor and the sunflowers scattered everywhere. With a quick jump Anina was in her bedroom. She slammed the door shut, grabbed her bed and pushed it up against the door. She had built a barricade that would keep her safe from the crocodile. Relieved, she let her breath out.

But then she hesitated. What if the beast was simply hungry? Maybe to make the crocodile go away you had to give it something to eat?

Anina looked again at the animal magazine. If the crocodile could crawl out of a picture then perhaps other animals could too. Anina hastily flipped through the magazine and stopped at a swarm of flamingos in a jungle swamp. Just right, she thought. They look like a birthday cake for crocodiles.

Suddenly there was a loud crack and the tip of the crocodile’s tail pushed through the splintered door.

Quickly, Anina held the picture of the flamingos up to the hole in the door and called as loud as she could, “Get out of the swamp! Shoo! Shoo!” Then she threw the magazine through the hole into the hallway, clapped her hands and yelled and screamed.
She could hardly believe what happened next. The entire hallway was suddenly filled with screeching flamingos wildly flapping their wings and running around all over the place on their long, skinny legs. Anina saw one bird with a sunflower in its beak and another grabbing her mother’s hat from its hook. She also saw a flamingo disappear into the crocodile’s mouth. With two quick bites he swallowed the flamingo and quickly followed it with another, the one with the sunflower in its beak.

After two portions of flamingo the crocodile seemed to have had enough and lay down contentedly in the middle of the hallway. When he had closed his eyes and no longer moved, Anina quietly opened her door and slipped through it into the hallway. She placed the empty magazine cover in front of the crocodile’s nose. “Please,” she whispered, “please go back home.” She crept back into the bedroom and looked through the hole in the door. She saw the crocodile back on the cover of the magazine.

She now went cautiously into the living room where the flamingos were crowded around the sofa and standing on the television. Anina opened the magazine to the page with the empty picture. “Thank you,” she said, “thank you very much. You may now go back to your swamp.”
In the morning, it was very difficult for her to explain the giant wet spot on the floor and the broken door to her parents. They weren’t convinced about the crocodile even though her mother’s hat was nowhere to be found.
1. What was the first sign that something unusual was happening?
   A. A pile of newspapers began to move.
   B. Anina saw the magazine cover.
   C. The door to her room was broken.
   * D. Anina heard a hissing sound.

2. Where did the crocodile come from?
   A. the bathroom
   * B. a magazine cover
   C. under the bed
   D. a nearby river

3. Which words tell you that Anina was frightened?
   * A. "frozen to the spot"
   B. "could not believe her eyes"
   C. "let her breath out"
   D. "sounded like a quiet hissing"

* Correct answer
4. Why did Anina think the crocodile was going to attack?

- It showed its long row of teeth.
- It let out a loud hissing sound.
- It started grunting and snorting.
- It swung its tail back and forth.

5. Put the following sentences in the order in which they happened in the story.

The first one has been done for you.

___ Anina saw the crocodile.
___ The crocodile ate two flamingos.
___ Anina tried to explain the broken door to her parents.
___ Anina started to walk to the bathroom.
___ Anina ran to the bedroom and slammed the door.

6. Why did Anina call the flamingos?

- Correct answer
7. How did the bedroom door get broken?
   A. The crocodile's tail pushed through it.
   B. The big vase cracked against it.
   C. The flamingo's sharp beak crashed into it.
   D. The bed smashed against it.

8. How did the magazine help Anina? Write two ways.
   1. ________________________________
      ________________________________
   2. ________________________________
      ________________________________

9. At the end of the story, how did Anina feel toward the flamingos?
   A. guilty
   B. cautious
   C. grateful
   D. annoyed

* Correct answer
10. Name one thing Anina had difficulty explaining to her parents.

__________________________________________________________________________

__________________________________________________________________________

11. You learn what Anina was like from the things she did. Describe what she was like and give two examples of what she did that show this.

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________
12. The author does not tell us whether Anina’s adventure was all a dream.

Give one piece of evidence that it may have been a dream.

Give one piece of evidence that it may not have been a dream.
Searching for Food

Here are three projects about the things small creatures eat and the ways they search for food. First you need to find actual ants, pill bugs, and worms. Treat them carefully and make sure you put them back where you found them after you have finished studying them.

- Follow an Ant Trail
- Study Pill Bugs
- Make a Wormery

Where to find ants, pill bugs, and worms

Ant trails are found in summer. At one end will be some food; at the other you should find the entrance to a nest.

Pill bugs like damp, dark places. They can be found under logs, under piles of dead leaves, and in walls.

Worms live under stones, in freshly dug soil or near compost heaps. They come to the surface at night.
Follow an Ant Trail

Ants live together in nests. When an ant finds some food it makes a trail for others to follow. To do this experiment you will need to find an ants' nest. You will also need the following materials: a sheet of paper, a small piece of apple, a handful of soil.

1. Put the piece of apple on the sheet of paper and lay the paper close to an ants' nest. Wait for some ants to find the apple. They should all follow the same trail.

2. Move the apple. Do the ants go straight to it?

3. Now sprinkle soil on the paper to cover the trail. The ants should scurry around for a while. Do they make a new trail?

What happens?
Even after the food has moved, the ants still follow the old trail until a new one is laid.

Why?
Once an ant has found some food, it produces special chemicals that leave a scent trail. Other ants from the nest use their antennae, or feelers, to sense this scent.
Study Pill Bugs

Pill bugs have sensitive antennae. Make this box, then collect six pill bugs in a container. Watch how they find their way when you put them in a box. You will need: a small empty box with a lid, scissors, adhesive tape, and dead, damp leaves.

1. Use the lid to make three long strips for making the passages in the picture.

2. Let your pill bugs walk along the passage one at a time. When they reach the end of the passage, some will turn left and some will turn right.

3. Put damp leaves in the right hand side of the box. Now let the pill bugs walk through the box again. Which way do they go?

What happens?
The pill bugs will turn to the right toward the food.

Why?
The pill bugs can sense the food with their antennae. They use them to find the leaves.
Make a Wormery

Worms are hard to study because they don’t like the light. As soon as they sense it, they wriggle away, trying to find a dark place again. To see how worms live and feed, make a wormery like the one shown here. Then find two or three worms to put in it. It is important to remember not to pull on the worms or you may hurt them. They are covered with bristles that grip the soil tightly.

You will need

- Shoe box
- Adhesive tape
- Pen
- Scissors
- Large plastic bottle
- 1 mug of sand
- 3 mugs of damp, crumbly soil
- Small cubes of onion and potato

1. Tape one side of the shoe box lid to the box, so it opens like a door. Poke holes in the top of the box with the pen to let air and light into the wormery.

2. Cut the top off the bottle. Then fill it with loosely packed layers of soil and sand. Scatter potato and onion on the surface.

3. Gently drop in your worms, then stand the bottle in the box and close the door. Leave it outside in a cool, dry place for four days.

4. After four days, go back and look at the bottle. What is different about the sand and soil?

Don’t forget: when you’ve finished with this project, put the worms back where you found them.
What happens?
After four days, the layers of sand and soil will have been mixed together.

Why?
The worms mix the sand and soil coming to the surface to eat the food and then tunneling underground to get away from the light.

From Animal watching in the Usborne Bug Book of Experiments published in 1996 by Usborne Publishing Ltd., London. An effort has been made to obtain copyright permission.
Questions: Searching for Food

1. What is the main purpose of the article?
   * A) to describe different projects you can do
   * B) to give information about ant trails
   * C) to show what small creatures look like
   * D) to explain what worms eat

2. What is one thing you should do to take care of the creatures?
   * A) search for them under rocks and stones
   * B) find out all about them
   * C) collect as many as you can
   * D) put them back where you found them

* Correct answer
Questions 3-5 are about the Ant Project

3. Why do you put the apple by the ants’ nest?
   - ☐ to block the ants’ trail
   - ✗ ☑ so the ants will make a trail
   - ☑ to confuse the ants
   - ☐ so the ants will scurry around

4. Once an ant finds some food, how do the other ants from the nest find it too?
   - ☐ They watch the first ant and follow it.
   - ☑ They run around until they find the food.
   - ✗ ☑ They sense the scent left by the first ant.
   - ☐ They smell the food on the piece of paper.

5. Why do the ants scurry around after you’ve sprinkled the soil?
   - ✗ Correct answer
6. How do pill bugs find the food?
   - A They walk down the passage.
   - ✫ B They sense food with their antennae.
   - C They follow the scent trail.
   - D They see the food in the dark.

7. Look at the picture for Study Pill Bugs. How does the picture help you to know what to do in the experiment?

   ✫ Correct answer
8. Why do you need to let your pill bugs walk along the passage before putting the leaves in the box?
   ○ To see if they can learn the maze.
   ✗ To see what they do when there is no food.
   ○ To see if the box is put together correctly.
   ○ To see which ones turn which way.

9. In Step 3 of the pill bugs project, what do you think will happen if you move the damp leaves to the left corner of the box?
   ✗

10. What is similar in the way ants and pill bugs find their food?
    ✗

* Correct answer
11. Number the steps in the order you would follow to make a wormery. The first one has been done for you.

   _____ put the bottle in the shoebox
   ____ poke holes in the top of the shoebox
   _____ drop in the worms
   _____ add potato and onion
   _____ fill the bottle with soil and sand

12. Explain why it is important to put layers of soil and sand in the bottle.

   (1) __________________________________________________________

   __________________________________________________________
13. Explain why putting the onion and potato on the surface of the soil is important to the wormery project.

____________________________________________________________________________________

____________________________________________________________________________________

____________________________________________________________________________________


14. Each project has What happens and Why in a separate box. What is the purpose of these boxes?

(A) to explain the steps of the project
(B) to tell you what you need for the project
(C) to tell you what to do when you have finished
(D) to explain what you have seen

* (D) to explain what you have seen

15. Which of the three projects did you find the most interesting? Use information from the text to explain your answer.

____________________________________________________________________________________

____________________________________________________________________________________

____________________________________________________________________________________

* Correct answer
Appendix B: Examining Class Composition Using a Multinomial Multilevel Logistic Regression Model

A multinomial multilevel logistic regression model was conducted using the HLM 6.08 software to investigate the composition of resulting latent classes (LCs). A multilevel model was used because of nesting of students within classrooms and its potential effect on the logistic regression model. As the DV (LC membership) was an unordered categorical variable, a multinomial model was used. The following four variables were included as level-1 (student-level) predictors: student achievement using average plausible values and student manifest characteristics including: students’ age, gender and country. Country was analyzed using a design matrix because it was a nominal variable with four categories. Qatar was used as the reference group leading to three comparison groups (Taipei versus Qatar, Hong Kong versus Qatar and Kuwait versus Qatar). These student manifest characteristics were selected because they are often used in manifest DIF analyses. All student-level predictors were group mean centered because they were of substantive interest (Enders & Tofighi, 2007). Level-2 (classroom-level) predictors were: the percentage of time students spent working with individual students or small groups, whether teachers taught students in same ability groups and whether teachers taught students strategies for decoding sounds and words. The latter two teacher-level variables were dummy coded variables and were not centered whereas the former teacher-level variable was continuous and was grand-mean centered (Peugh, 2010).

To investigate whether there was a need for multilevel modelling, the intra-class correlation (ICC) for (ordered or unordered) categorical DVs (Hedeker, 2003) and the design effect statistics (DEFF) (Peugh, 2010) were calculated. The ICC is the proportion of variance at the classroom-level (Level-2) once level-1 (student-level) variance has been taken into account (Peugh, 2010). The DEFF takes the cluster size into account and indicates how much the standard errors are underestimated in a complex versus a simple random sample (Maas & Hox, 2005). It is important to note that
although PIRLS 2006 sampled intact classrooms in each school, background questionnaire data associated with examinees’ item responses for the Reader booklet consisted on average of a small number of examinees (approximately 3 to 5 students) per classroom; therefore, the cluster sizes investigated herein were fairly small. Previous research suggests that a DEFF value below 2 indicates there is no need to run a multilevel model (Muthén & Satorra, 1995).

The following model was run using LC membership as a DV:

\[
\eta_{ij} = \gamma_{00} + \gamma_{01} * \text{ABGRP}_j + \gamma_{02} * (\overline{\text{SGRP}}_j - \overline{\text{SGRP}}) + \gamma_{03} * \text{DECOD}_j + \gamma_{10} * (\overline{\text{GEN}}_i - \overline{\text{GEN}}) + \gamma_{20} * (\overline{\text{AGE}}_j - \overline{\text{AGE}}) + \gamma_{30} * (\overline{\text{TQA}}_j - \overline{\text{TQA}}) + \gamma_{40} * (\overline{\text{KQA}}_j - \overline{\text{KQA}}) + \gamma_{50} * (\overline{\text{HQA}}_j - \overline{\text{HQA}}) + \gamma_{60} * (\overline{\text{PVAL}}_j - \overline{\text{PVAL}}) + \mu_{0j}
\]

In this model, level-2 predictors: ABGRP represented how often teachers spent organizing students in same ability groups when doing reading activities, SGRP was the percentage time teachers spent in class teaching individual students or small groups, DECOD indicated how often teachers taught students strategies for decoding sounds and words. Level-1 predictors: AGE represented students’ age, TQA was the comparison of Taipei versus Qatar, KQA was the comparison of Kuwait versus Qatar, HQA compared Hong Kong versus Qatar and PVAL was student average plausible value. The numbers of Level-1 and Level-2 predictors as well as the average number of examinees per classroom are summarized in Table B.1.

Table B.1

<table>
<thead>
<tr>
<th></th>
<th>LC1 vs. LC2</th>
<th>LC1 vs. LC3</th>
<th>LC2 vs. LC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-1</td>
<td>1451</td>
<td>1913</td>
<td>1908</td>
</tr>
<tr>
<td>Level-2</td>
<td>449</td>
<td>428</td>
<td>467</td>
</tr>
<tr>
<td>Students per Class</td>
<td>3.23</td>
<td>4.47</td>
<td>4.09</td>
</tr>
</tbody>
</table>
Cases with missing data were deleted from the analyses. Total sample sizes were 728 examinees in LC1, 723 examinees in LC2 and 1185 examinees in LC3. Table B.2 summarizes results of the multilevel multinomial logistic regression analyses. These results revealed ICC values ranging from 0.21 to 0.56, which indicated that classroom-level variance accounted for 21% to 56% of variance in LC membership depending on which LCs were compared. DEFF values ranged from 1.55 to 2.73, indicating the need for a multilevel model analysis for one comparison group (LC2 versus LC3) based on DEFF > 2 (Peugh, 2010). Because of nesting for at least one of the comparison groups, a multilevel model was run for all comparison groups. As shown in Table B.2, average plausible value was a significant student-level predictor ($p < .01$) across all comparison groups. All other student-level predictors were not significant except country (Qatar versus Hong Kong) for LC1 versus LC3. All three examined teacher-level predictors were significant ($p < .01$) for LC1 versus LC3 and LC2 versus LC3 whereas teaching of same ability groups was the only significant teacher-level predictor for LC1 versus LC2. Associated odds ratio results for the variables that were found to be significant are also summarized in Table B.2.
Table B.2

Results of Multinomial Multilevel Analyses

<table>
<thead>
<tr>
<th></th>
<th>LC1 vs. 2</th>
<th>Odds</th>
<th>LC 1 vs. 3</th>
<th>Odds</th>
<th>LC 2 vs. 3</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>.25</td>
<td>N/A</td>
<td>.21</td>
<td>N/A</td>
<td>.56</td>
<td>N/A</td>
</tr>
<tr>
<td>DEFF</td>
<td>1.55</td>
<td>N/A</td>
<td>1.74</td>
<td>N/A</td>
<td>2.73</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Student-Level Predictors

<table>
<thead>
<tr>
<th></th>
<th>LC1 vs. 2</th>
<th>Odds</th>
<th>LC 1 vs. 3</th>
<th>Odds</th>
<th>LC 2 vs. 3</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plausible values</td>
<td>&lt;.01***</td>
<td>1.021</td>
<td>&lt;.01***</td>
<td>.977</td>
<td>&lt;.01***</td>
<td>.988</td>
</tr>
<tr>
<td>Age</td>
<td>.245</td>
<td>NS</td>
<td>.034</td>
<td>NS</td>
<td>.542</td>
<td>NS</td>
</tr>
<tr>
<td>Gender</td>
<td>.071</td>
<td>NS</td>
<td>.695</td>
<td>NS</td>
<td>.658</td>
<td>NS</td>
</tr>
<tr>
<td>Taipei vs. Qatar</td>
<td>.547</td>
<td>NS</td>
<td>MC</td>
<td>NS</td>
<td>.822</td>
<td>NS</td>
</tr>
<tr>
<td>Hong Kong vs. Qatar</td>
<td>.012</td>
<td>NS</td>
<td>&lt;.01***</td>
<td>.024</td>
<td>.841</td>
<td>NS</td>
</tr>
<tr>
<td>Kuwait vs. Qatar</td>
<td>MC</td>
<td>NS</td>
<td>MC</td>
<td>NS</td>
<td>MC</td>
<td>NS</td>
</tr>
</tbody>
</table>

Teacher-Level Predictors

<table>
<thead>
<tr>
<th></th>
<th>LC1 vs. 2</th>
<th>Odds</th>
<th>LC 1 vs. 3</th>
<th>Odds</th>
<th>LC 2 vs. 3</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small group teaching</td>
<td>.428</td>
<td>NS</td>
<td>&lt;.01***</td>
<td>1.05</td>
<td>&lt;.01***</td>
<td>1.054</td>
</tr>
<tr>
<td>Same ability groups</td>
<td>&lt;.01***</td>
<td>2.311</td>
<td>&lt;.01***</td>
<td>.484</td>
<td>&lt;.01***</td>
<td>.248</td>
</tr>
<tr>
<td>Decoding</td>
<td>.407</td>
<td>NS</td>
<td>&lt;.01***</td>
<td>1.365</td>
<td>&lt;.01***</td>
<td>1.316</td>
</tr>
</tbody>
</table>

Note: ICC=Intra-Class Correlation
DEFF = Design Effect Statistic
LC=Latent Class
*** p<.01
NS= Not Significant
N/A = Not Applicable

As mentioned, plausible values as a predictor variable was significant across all LC comparisons. The odds ranged from .977 to 1.021, which indicated that an increase in plausible values led to substantially higher odds of being in a higher performing LC. The comparison between
Hong Kong versus Qatar was also significant, the odds however were fairly small (.024) which indicated that being from Hong Kong led to slightly higher odds of being in a higher performing LC (LC3), in addition to plausible values as an independent variable. The variable SGRP (teaching in small groups) was significant across (LC1 vs. 3 and LC2 vs. 3) which indicated that the greater percentage of time teachers spent working with individual students, the higher the odds examinees would be in the higher achieving LC3. The odds ratio results (1.05) indicated a substantial amount of variation associated with teaching students in small groups or individually. Results of a significant ABGRP variable suggest that the more time teachers spend on working with same ability groups, the higher the odds of being in a higher achieving LC. Although, the odds were small when comparing LC1 versus LC3 (.48) and LC2 versus LC3 (.248), the odds more than doubled when comparing LC1 versus LC2 (2.311). The decoding variable was also significant which indicated that examinees who were taught by teachers who spent less time on teaching decoding abilities, had higher odds of being in a higher performing LC. One possible reason is that this may liberate teachers’ time to spend on other activities such as teaching more complex reading strategies.