SCORE SCALE COMPARABILITY IN
INTERNATIONAL EDUCATIONAL ASSESSMENTS

by

Debra Anne Sandilands

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

MASTER OF ARTS

in

The Faculty of Graduate Studies
(Measurement, Evaluation and Research Methodology)

University of British Columbia

(Vancouver)

December 2008

©Debra Anne Sandilands, 2008
ABSTRACT

Many countries, including Canada, are increasingly using international educational assessments to make comparisons of achievement across countries and to make important decisions regarding issues such as educational policy and curriculum. Most large-scale assessments have different forms that are adapted and/or translated for use across multiple language and cultural groups. Equivalence and fairness for examinees of all groups must be established in order to support valid score comparisons across groups and validity of decisions made based on these assessments.

This study investigated the degree of score comparability in the Reader booklet of the Progress in International Reading Literacy Study (PIRLS) 2001 at three levels of scores. At the item level, differential item functioning (DIF) analyses were conducted using Ordinal Logistic Regression and Poly-SIBTEST. DIF items were grouped into bundles and analyzed for differential bundle functioning (DBF) using Poly-SIBTEST. Differences in item response theory-based test characteristic curves (TCCs) were analyzed to investigate comparability at the scale level. The study focussed on four countries: Argentina, Colombia, England and USA.

The results of this study confirm previous studies that demonstrate a large degree of DIF in international educational assessments. Results also indicate a high degree of similarity between the two DIF methods used in identifying DIF items, but fail to support the correspondence between their effect size measures. This study expands on the research base regarding DBF and demonstrates a two stage approach to identifying potential causes of differential functioning. Results of DBF analyses indicate that cognitive levels tapped by reading comprehension questions may represent a source of bias leading to differential functioning in the Reader booklet. This study also contributes preliminary evidence for the possibility that the use of international item parameters to create individual country scores may provide a relative advantage to some countries due to the locations of their score distributions, which may have implications regarding current score scale creation methods.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF ACRONYMS</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Overview</td>
<td>1</td>
</tr>
<tr>
<td>Bias and Factors that Affect Score Scale Comparability</td>
<td>3</td>
</tr>
<tr>
<td>Current Scaling Methods in Large-scale Assessment</td>
<td>7</td>
</tr>
<tr>
<td>Purpose of Study and Research Questions</td>
<td>11</td>
</tr>
<tr>
<td>CHAPTER II: LITERATURE REVIEW</td>
<td>13</td>
</tr>
<tr>
<td>Large-scale Assessment in the Canadian Context</td>
<td>13</td>
</tr>
<tr>
<td>Statistical Methods for Establishing Score Comparability</td>
<td>16</td>
</tr>
<tr>
<td>Establishing Construct Comparability: Analysis of Factor Structures</td>
<td>17</td>
</tr>
<tr>
<td>Establishing Item Level Equivalence: Differential Item Functioning Analyses</td>
<td>18</td>
</tr>
<tr>
<td>Ordinal Logistic Regression</td>
<td>20</td>
</tr>
<tr>
<td>Poly-SIBTEST</td>
<td>21</td>
</tr>
<tr>
<td>Establishing Bundle Level Equivalence: Differential Bundle Functioning Analyses</td>
<td>25</td>
</tr>
<tr>
<td>Establishing Test Level Equivalence: Comparison of TCCs</td>
<td>28</td>
</tr>
<tr>
<td>Summary</td>
<td>30</td>
</tr>
<tr>
<td>CHAPTER III: METHOD</td>
<td>31</td>
</tr>
<tr>
<td>Overview</td>
<td>31</td>
</tr>
<tr>
<td>Measure</td>
<td>31</td>
</tr>
<tr>
<td>Data</td>
<td>41</td>
</tr>
<tr>
<td>Analyses</td>
<td>45</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: PIRLS 2001 Aspects of Reading ........................................ 34
Table 2: PIRLS 2001 Distribution of Literary and Informational Blocks
   Across Booklets................................................................. 36
Table 3: PIRLS 2001 Sample Information and Overall Average Scale
   Scores.................................................................................. 39
Table 4: PIRLS 2001 Reader Booklet Sample Sizes............................ 42
Table 5: PIRLS 2001 Reader Booklet Items...................................... 44
Table 6: PIRLS 2001 Reader Booklet Descriptive Statistics............... 59
Table 7: Ordinal Logistic Regression DIF Results.............................. 61
Table 8: Poly-SIBTEST DIF Results............................................ 62
Table 9: Number of DIF items identified by effect size by method........ 63
Table 10: Items Identified as DIF by Ordinal Logistic Regression and
   Poly-SIBTEST..................................................................... 64
Table 11: Differential Bundle Functioning Analyses.......................... 67
Table 12: Average original and transformed discrimination and location
   parameters............................................................................ 69
Table 13: Correlation of a and b parameters for each country with the
   international parameters....................................................... 70
Table 14: Descriptive statistics of differences between country TCCs and
   international TCCs............................................................... 73
Table 15: Q1 Goodness of fit results.............................................. 76
Table 16: PCA Eigenvalues and variance explained for each factor........ 77
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Argentina and Colombia Score Distributions</td>
<td>60</td>
</tr>
<tr>
<td>Figure 2</td>
<td>England and USA Score Distributions</td>
<td>60</td>
</tr>
<tr>
<td>Figure 3</td>
<td>TCC Comparison for Argentina</td>
<td>71</td>
</tr>
<tr>
<td>Figure 4</td>
<td>TCC Comparison for Colombia</td>
<td>71</td>
</tr>
<tr>
<td>Figure 5</td>
<td>TCC comparison for England</td>
<td>72</td>
</tr>
<tr>
<td>Figure 6</td>
<td>TCC comparison for USA</td>
<td>72</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Argentina and Colombia, Differences in TCCs based on international versus country parameters</td>
<td>74</td>
</tr>
<tr>
<td>Figure 8</td>
<td>England and USA, Differences in TCCs based on international versus country parameters</td>
<td>74</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Scree plots for Reader booklet data</td>
<td>78</td>
</tr>
</tbody>
</table>
LIST OF ACRONYMS

1PL One parameter logistic model
2PL Two parameter logistic model
2PPC Two parameter partial credit model
3PL Three parameter logistic model
AERA American Educational Research Association
APA American Psychological Association
DBF Differential bundle functioning
DFIT Differential Fit of Items and Tests
DIF Differential item functioning
DTF Differential test functioning
EFA Exploratory factor analysis
GMH Generalized Mantel-Haenszel
GPCM Generalized Partial Credit Model
GRM Graded Response Model
ICC Item characteristic curve
IEA International Association for the Evaluation of Educational Achievement
IRT Item response theory
ITC International Test Commission
LDFA Logistic discriminant function analysis
MH Mantel-Haenszel
MMD Multi-dimensional model for DIF
NCME National Council on Measurement in Education
LR Logistic Regression
OEL Ordering of the expected latent trait
OLR Ordinal logistic regression
PIRLS Progress in International Reading Literacy Study
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PISA</td>
<td>Program for International Student Assessments</td>
</tr>
<tr>
<td>Poly-SIBTEST</td>
<td>Polytomous SIBTEST</td>
</tr>
<tr>
<td>SIBTEST</td>
<td>Simultaneous item bias test</td>
</tr>
<tr>
<td>SMD</td>
<td>Standardized Mean Difference</td>
</tr>
<tr>
<td>TCC</td>
<td>Test characteristic curve</td>
</tr>
<tr>
<td>TIMSS</td>
<td>Trends in International Mathematics and Science Study</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to express sincere gratitude to my thesis supervisor, Dr. Kadriye Ercikan, for her invaluable guidance over the past few years. I want to thank her not only for her thesis advice, but also for including me in many concurrent projects which have greatly broadened my experience and knowledge. I have benefitted tremendously from the highest quality of graduate education under Dr. Ercikan’s mentorship.

I am deeply indebted to my committee members, Dr. Nancy Perry, Dr. Jenna Shapka, and Dr. Marion Porath for their constructive feedback and assistance. I am especially grateful to each of them for their willingness to work on a very tight time frame to enable me to achieve this goal.

My dear friend and fellow student, Malena Oliveri, has been an amazing source of inspiration and support to me along this journey. What started as a casual friendship in a class on item response theory has become one of the more meaningful friendships I have had the privilege of enjoying.

I also want to thank my family for their confidence in me and their never-ending support. My daughters, Kate and Stephanie, and my husband, Andy, have been pillars of strength for me to lean on when needed. Their love has lifted me and carried me along.
CHAPTER 1: INTRODUCTION

Overview

Many countries, including Canada, are increasingly using international, large-scale educational assessments to address issues such as achievement and accountability (Crundwell, 2005). Currently, assessments such as Progress in International Reading Literacy Study (PIRLS), the Program for International Student Assessments (PISA) and Trends in International Mathematics and Science Study (TIMSS) are administered in as many as 57 countries simultaneously in dozens of different languages. These international assessments are used to make comparisons of educational achievement across countries or regions, decisions regarding educational policy, curriculum, program development and evaluation. Since their uses and inferences are concerned with making comparisons, issues regarding appropriate and justifiable comparability of results from these assessments are critical to consider. According to the Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999) (the “Standards”), comparisons across groups are only meaningful if scores are comparable across groups. Establishing score comparability is therefore critical to establishing validity of group comparisons in international assessments.

The need to establish comparability increases as the stakes resulting from the assessments increase. When used for measuring systems for policy and accountability purposes, large-scale international assessments are high-stakes
assessments (dePascale, n.d.). Therefore it is particularly important to establish that the scores for different groups of test takers are comparable.

International assessments typically involve testing a wide variety of examinees with one assessment tool that is intended to be comparable for all groups of examinees. Most large-scale assessments are constructed to have different but comparable forms, which can be used across multiple language and cultural groups. They cannot be assumed to be equivalent and free of bias for all examinees of different cultural, ethnic, language and curricular backgrounds. This equivalence and fairness must be established through the use of quantitative empirical methods.

Establishing equal validity of score interpretations for all examinees is a complicated and essential task for test developers. Both the International Test Commission Guidelines (International Test Commission, 2000) (the “ITC Guidelines”) and the Standards (AERA, APA & NCME, 1999) place a responsibility on test developers to provide evidence that all language versions of tests are equivalent in order to support score comparability. However, currently, test developers are not providing this critical evidence to support score comparability.

Equivalence of test scores across test versions requires that the different test versions measure the same construct with the same level of measurement accuracy. Much research has indicated that these assumptions are not tenable in large-scale assessments. If this is the case, test developers should use “comparable scales” which result in scalar equivalence (Cook, 2006) rather than
using equated scores. For scalar equivalence, scores must have the same measurement unit and origin in all populations and the measurement must be free of bias (van de Vijver & Tanzer, 2004).

There are many sources of bias that jeopardize comparability in large-scale international assessments such as Progress in International Reading Literacy Study 2001 (“PIRLS 2001”). They must be investigated as a prerequisite to determining scalar equivalence. The next section will provide a more comprehensive definition of bias and a discussion of the factors that have been shown to introduce bias and reduce score scale comparability. It begins with the presentation of three conceptualizations of bias and its sources, and ends by noting the common features between these interpretations, which will encapsulate the definition of bias taken throughout this current research.

Bias and Factors that Affect Score Scale Comparability

In the Standards (AERA, APA & NCME, 1999), bias is conceptualized as group differences that result from deficiencies in a test itself or the manner in which it is used rather than true differences in ability on the construct being measured, where a construct is “the concept or characteristic that a test is designed to measure.” The Standards define bias as construct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers. There are two general sources of bias discussed in the Standards: bias associated with test content and bias associated with response processes.
Bias related to test content can be introduced when there is an inappropriate sampling of test content, for example, when the content of the test does not adequately cover the intended construct, is not relevant to the intended construct, or is not aligned with the curricula for all groups. Occasionally, testing material may be differentially interesting or motivating to different groups, which results in bias related to the test content. Response-related sources of test bias include the use of scoring criteria that may not adequately represent different response approaches taken by different groups. Test items can elicit responses other than those predicted or intended by test developers. The use of particular response formats may cause impediments for some groups or advantages for others depending on familiarity or differential abilities with the response format.

Bias has also been conceptualized as one of two broad aspects of fairness, one of which focuses on the validity of an assessment. Bond, Moss, and Carr (1996) define bias as an interpretation or action based on a test that is not equally valid for all relevant subgroups. In other words, bias and validity share as their point of reference the interpretations, uses or decisions that flow from an assessment, rather than the assessment itself. Bond et al. characterize bias in two ways, one being the situation where scores systematically under-estimate the performance of one group on the construct intended to be assessed. The other occurs when an assessment is intended to measure the same construct across the groups, but actually measures different constructs in the groups. Sources of bias can be internal to the assessment, such as bias that arises from the specification of the content framework, the training of scorers, the development of the tasks and
exercises being assessed, and the manner in which they are scored. External sources of bias can arise when the behaviours elicited from the test are not those the test is intended to predict or when the conditions of testing are not related to those under which the construct would normally be exhibited.

Van de Vijver and Tanzer (2004) also provide a thorough discussion of bias and the sources of bias which affect scalar equivalence. Their conceptualization of bias and its sources is summarized in the following paragraphs.

Bias refers to nuisance factors which are present in an assessment that lead to score differences for one group which do not reflect or correspond to differences in the underlying trait which is being measured. Equivalence is defined as a lack of bias, and is required for valid comparisons across cultural or other groups. Three kinds of bias that affect scalar equivalence are identified in the van de Vijver and Tanzer (2004) overview.

The first type of bias distinguished is construct bias. Construct bias occurs when the construct being measured is not the same for all groups or when the behaviours associated with the construct are differentially appropriate in the groups.

Method bias arises from the administration of the instrument, or from the instrument itself. Method bias has significant consequences on validity as it leads to shifts in average scores and affects differences observed in scores at the test level. Method bias is categorized into three types: (1) sample bias, which leads to incomparability of samples due to aspects not related to the target variable such as
educational background or motivation levels of examinees; (2) instrument bias, which refers to the fact that different groups respond in different manners to the form of an assessment or to the type of response required by the assessment; and (3) administration bias, which arises from problems or differences across administration of the assessment which compromise the collection of data.

The third source of bias, which affects scalar equivalence noted by van de Vijver and Tanzer (2004) is termed item bias and is the only one of the three types of bias which occurs at the item level, that is, at the level of the individual question. Biased items have different meaning across groups. Item bias is one form of differential item functioning (DIF), which occurs when persons from different groups, matched on ability, do not have the same probability of correctly responding to, or endorsing, an item (Hambleton, Swaminathan, & Rogers, 1991). In other words, they will not achieve the same expected score on an item. Item bias is differentiated from item impact, which occurs when two groups have a different probability of correctly responding to an item as a result of true ability differences between the groups on the construct of interest. While item impact is related to group characteristics, item bias is a result of an interaction between group characteristics and item characteristics. The finding of empirical evidence of DIF is necessary but not sufficient to conclude bias is present. Once an item has been determined to exhibit DIF, it is necessary to use judgment to decide whether the item is biased.

While slight differences can be seen in the interpretations of bias presented above, it is also clear that there are common foundations or overarching themes
among them. First, bias is akin to "differential invalidity" related to group differences in scores and in score interpretations that are not related to the construct or facet of ability intended to be measured. In other words, when bias exists, student scores are not only a function of their ability on the construct of interest; in addition, there is some degree of variance contributing to scores that is not accounted for by ability and this results in a lack of score comparability on the construct of interest. Second, bias can be introduced into an assessment at all stages, beginning with the test development stage where the construct, content, curriculum, format, translation and adaptation, response processes and scoring schemes are critical to consider for each potential subgroup in the population. At the administration stage, bias can be introduced through sampling methodology and inconsistencies in administration that compromise the collection of data. At the scoring stage, bias can result from inconsistencies in training of scorers or in scorers' applications of scoring procedures.

Current Scaling Methods in Large-scale Assessment

In most large-scale assessments, current scaling methods use Item Response Theory ("IRT") to create item parameters. In the 3-Parameter Logistic IRT Model ("3PL model"), the three item parameters - item difficulty, discrimination, and pseudo-guessing - are estimated in addition to the ability or latent trait of an examinee. According to IRT, the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an Item Characteristic Curve ("ICC") which specifies that as the level of the trait (i.e. a
latent trait or ability) increases the probability of a correct response to an item increases (Hambleton et al. 1991). The sum of all the item characteristic curves in a test gives a Test Characteristic Curve ("TCC") for a given assessment.

Current scaling methods in international assessments administered in multiple languages are based on an international sample which is created using samples of data from all countries, with the country samples being weighted so that each country contributes equally to the overall international sample. The international sample is intended to be representative of all of the individual countries. Score scale comparability is directly related to the appropriateness of the international parameters for individual countries. The appropriateness of item parameters derived from an international sample for individual countries is necessary but not sufficient for score creation methods that support comparability. International item parameters must fit the country data well in order to create score scales and make comparisons across countries (Ercikan & Sandilands, in press; Ercikan & Gonzalez, 2008). If the international item parameters do not fit country data well, then TCCs based on international calibration and those based on country calibrations would be expected to have differences and lead to score incomparability. Since poor item fit can accumulate across items, it is possible for overall poor fit to be hidden by the fact that some items may favor a country while others disfavor it, thereby cancelling each other out. In this case there would be no apparent difference at the test level when analyzed using international parameters or using individual country parameters. It might still be possible for
scalar incomparability to exist without being evident in TCCs or in score
distribution comparisons.

At the item level, previous research has indicated that there is a large
degree of DIF in large-scale assessments that have been translated. Ercikan, Gierl,
McCreith, Puhan, and Koh (2004) studied bilingual versions of Canada’s national
achievement tests. Their results point to substantial psychometric differences
between language versions at the item level with up to 36% of the items being
identified as DIF for the 2 language groups. Sireci and Allalouf (2003) found that
34% of the items on Israel’s Psychometric Entrance Test exhibited DIF between
the Russian and Hebrew versions.

Despite large amounts of DIF in large-scale assessments, preliminary
research in comparability at the test level seems to indicate that few to no
differences are seen at the test level. In a study of the potential for manifestation
of item level DIF at the test (scale-level), Zumbo (2003) used both exploratory and
confirmatory factor analysis as well as an analysis of reliability. He found that the
presence of DIF was not necessarily reflected at the whole test level. Ercikan and
Sandilands (in press) investigated the TIMSS 2003 grade 8 mathematics
assessment and found that there may be only a small (i.e., 5% of maximum score)
difference between the international and country TCCs. Ercikan and Gonzalez
(2008) conducted a study in which they compared country TCCs to the
international TCC using the PIRLS 2001 data. Their study also indicated that
there was very little difference between them, with the largest score scale
difference being 5% of the maximum possible score. These studies and others
(Drasgow, 1987; Pae, 2004) seem to lend support to the use of international parameters to create country scores. However, while small, a 5% difference might have a significant impact if many examinees are found in the score ranges where the differences are seen.

In order to gain an understanding of this apparent discrepancy between comparability at the item level (large amounts of DIF items found) and at the test level (small differences in test characteristic curves) in large-scale assessments, the current research will extend the investigation of comparability by examining DIF amplification or cancellation at the bundle level. A bundle is a group of test items that is thought to share, and measure, a common factor or dimension that may be a source of differential functioning. Statistical methods for identifying bundles of items as sources of differential functioning are more powerful than methods for identifying single items (Gierl, Bisanz, Bisanz & Boughton, 2001). Nandakumar (1993) demonstrated that the cumulative effect of DIF expressed through a bundle of items could lead to amplification, partial cancellation, or complete cancellation of DIF at the test level. She argued that even when a test contains individual items that function differently across subgroups, the test as a whole may not, due to partial or complete cancellation of DIF. The current research will examine bundles of items to explore the possibility of DIF cancellation being responsible for the apparent inconsistency between results of item level analyses and test level analyses in large-scale assessment.
Purpose of Study and Research Questions

The focus of this study is score scale comparability in international assessments. Specifically, using data from the PIRLS 2001 "Reader Booklet", this study investigated whether there is equivalence between score scales for countries at the item level (DIF), at the bundle level (differential bundle functioning, DBF), and at the test level (differential test functioning, DTF).

In addition, this study investigated whether different results are obtained when using different statistical methods to examine comparability. At the item level, the PIRLS data was examined for DIF using two methods: Ordinal Logistic Regression (Zumbo 1999) and Poly-SIBTEST (Chang, Mazzeo, & Roussos, 1996; Shealy & Stout, 1993). Since Poly-SIBTEST can also be used to detect differential functioning at the bundle level, it was employed to examine potential differential bundle functioning (DBF). At the test level, a comparison of test characteristic curves resulting from the use of international parameters versus individual country parameters was conducted.

This study leads to a greater understanding of the issues in comparability and validity in large-scale assessment, specifically about methodology of how scores and scales are created, and whether or not current methods support the uses, inferences, and comparison practices among the countries that participate in international assessment.

In summary, this research seeks to provide answers to the following four research questions:
1. What is the degree of DIF in the PIRLS data? What are similarities in DIF detection by OLR and Poly-SIBTEST?

2. At the bundle level, are there any bundles of items in the PIRLS Reader booklet that may result in DIF amplification or cancellation?

3. At the test level, what are similarities or differences in test characteristic curves based on the international parameters versus those based on individual country parameters? Are the international parameters appropriate for creating country scores?

4. Why do large degrees of differential functioning in large-scale assessments at the item level result in only small degrees of differences at the test level?
CHAPTER II: LITERATURE REVIEW

This chapter provides a review of current uses of large-scale assessments, particularly in the Canadian context, thereby affirming the need to address issues of validity especially related to establishing score comparability. Requirements placed on test developers to provide evidence of score comparability by the Standards (AERA, APA & NCME, 1999) and ITC Guidelines (ITC, 2000) are briefly reviewed. The remaining sections give a review of the types of evidence that should be provided in order to support score comparability, as well as a review of and justification for the use of the statistical methods for establishing score comparability.

Large-scale Assessment in the Canadian Context

Most Canadian provinces and territories conduct some form of large-scale assessment and use the results for holding educational systems accountable and for evaluating the effectiveness of school programs and systems (Volante, 2006). The British Columbia Ministry of Education uses large-scale provincial, national and international assessments to obtain information for the purposes of public accountability, to improve programs and schools, determine students’ knowledge and skills, and inform decisions to be made at the classroom, district and provincial levels for improving student achievement (British Columbia Ministry of Education, 2004).

In the Province of Ontario, the Educational Quality and Accountability Office, an agency of the Ontario provincial government whose mandate is to
provide parents, teachers and the public with accurate and reliable information about student achievement, reports that the findings from large-scale assessments are used for setting educational priorities and for improvement planning as well as to measure Ontario students’ achievement against national and international benchmarks (Education Quality and Accountability Office, 2008).

One large-scale assessment used in Canada at the international level is Progress in Reading Literacy Study (PIRLS). The Canadian Provinces of Ontario and Quebec took part in the 2001 administration of PIRLS. In the 2006 administration, Alberta, British Columbia, Nova Scotia, Ontario and Quebec participated. A 2007 news release from the British Columbia Ministry of Education regarding that province’s results in the 2006 PIRLS assessment exemplifies the extent to which results from this assessment are used for policy and program planning in British Columbia:

“B.C. is a world leader in literacy and this international assessment proves that. … The results we see here show that our ReadNow BC literacy strategy is working to make our province the most literate jurisdiction, not only in North America, but in the world. …Since 2001, the government has invested over $125 million in literacy for British Columbians through the Province’s literacy strategy, ReadNow BC.” (British Columbia Ministry of Education, 2007).

This news release exemplifies that the Province of British Columbia has relied on a comparison of its PIRLS 2006 scores with those of other jurisdictions and
countries, and interpreted a favourable comparison as justification for the financial
ccommitment of over $125 million and for the continued existence of the ReadNow
BC literacy strategy.

A similar news release by Alberta’s Ministry of Education revealed the use
of PIRLS 2006 results in that province to make comparisons between itself and the
rest of the world with the headline, “Alberta Grade 4 students achieved top three
in the world in the latest international literacy tests”. This was followed by
statements attributing this success in PIRLS to the quality of teaching, resources,
curriculum and programs in Alberta, and a justification for continued participation
in international large-scale testing as a means of demonstrating the success of
Alberta’s students in relation to the rest of the world (Alberta Education, 2007).

Clearly, large-scale assessments including PIRLS are employed for
comparisons and high stakes decision making by the individual countries and
jurisdictions that take part in them, as is manifested in the Canadian context. Yet,
what evidence is there that such interpretations of score comparisons are
meaningful and valid? What evidence should be provided?

According to the Standards (AERA, APA & NCME, 1999), comparisons
across groups are only meaningful if scores have comparable meaning across
groups. Standard 9.9, in particular, requires that test developers report evidence of
test comparability when different language versions are intended to be
comparable, as is the case in large-scale assessments such as PIRLS. Standard 7.3
puts an onus on test developers to conduct research to detect and eliminate aspects
of test design, content and format that might lead to DIF. The ITC Guidelines
caution against interpreting score differences between populations as though they represent true differences in the construct measured by the test and that interpretations be limited to those for which satisfactory validity evidence has been presented. Specifically, *ITC Guidelines* D.2, D.6, D.7, D.8, and D.9 place an onus on test developers and publishers to apply appropriate statistical techniques to establish equivalence, and to provide validity evidence of adapted test versions for all intended populations. In short, both the *Standards* and the *ITC Guidelines* require that the validity of interpretations of assessment results for all relevant groups of examinees be established through the use of empirical and qualitative methods. The focus of the current study is on statistical methods for establishing score comparability.

Statistical Methods for Establishing Score Comparability

There are many factors in large-scale assessments that can contribute to bias and threaten score equivalence. It is generally accepted that international large-scale assessments are not bias-free as a result of the adaptation and translation processes that can introduce bias and cause difficulties in comparability and equivalence (Allalouf, Hambleton, & Sireci, 1999; Ercikan et al, 1998; Ercikan, 2002; Gierl & Khaliq, 2001; Hambleton, 1993). Any analysis of score equivalence therefore must assess for bias, and weigh the evidence as to the degree of score comparability that exists.

The choice of statistical methods for establishing score comparability depends in part on the scoring schemes of the items being analyzed. The PIRLS
2001 Reader Booklet, like many other large-scale international assessments, contains two types of item scores, dichotomous and polytomous. Dichotomously-scored items provide binary data which is scored in only one of two levels, for example, correct/incorrect, or true/false. Polytomously-scored items, also referred to as partial credit or graded response items, provide ordinal data which can be scored in more than two levels, such as on a scale of 0 to 3. The PIRLS Reader contains multiple choice items which were scored dichotomously, as well as constructed response items which were scored polytomously, with up to 4 possible score levels. Given this, any statistical method considered for inclusion in this study or in any other similar study of large-scale assessments containing both types of items must be able to accommodate both dichotomously and polytomously scored data.

*Establishing Construct Comparability: Analysis of Factor Structures*

Bias can be introduced into an assessment when an assessment, intended to measure the same construct across groups, measures different constructs in the groups (Bond et al., 1996). Validity of group comparisons depends, in part, on whether construct comparability exists across the groups.

In addition, some psychometric methods such as IRT which were used in this study to examine comparability at the test level and which is used in current scaling methods in large-scale assessments have strict assumptions that must be met before they can be properly applied. One critical assumption is that there is a single latent trait that underlies test performance and that it is the same unidimensional construct across groups. An examination of unidimensionality of
the construct across the groups provided evidence as to whether this IRT assumption is met and hence the suitability of the use of IRT for this particular study. It also provided information regarding the appropriateness of PIRLS 2001 use of IRT to generate item parameters as the basis for establishing scores.

This study used exploratory factor analysis (EFA) to examine construct comparability. EFA is commonly used in studies of comparability of different language versions of large-scale academic assessments (Puhan & Gierl, 2006; Robin, Sireci, & Hambleton, 2003).

Establishing Item Level Equivalence: Differential Item Functioning Analyses

DIF analyses are used to examine item level equivalence across groups. DIF, as defined by Hambleton et al. (1991), occurs when individuals having the same ability, but from different groups, do not have the same probability of correctly responding to an item. The presence of DIF is related to group membership (such as language, culture, gender or country of residence) which in theory should not play a role in determining an examinee’s score on an individual item or on an assessment as a whole. This study uses the common terminology for naming groups in DIF studies, that is the “reference group” which serves as a group for comparison purposes, and the “focal group” which is the group of interest to the researcher.

There are several methods for investigating DIF. They are grouped into two main categories, those that are parametric and those that are nonparametric. Parametric methods assume that a specific mathematical model underlies the data. Of critical practical importance is that in order to use parametric analyses for DIF,
the data must be shown to fit the statistical model used. In contrast, nonparametric methods make no statistical assumptions and there is no resulting need to establish fit between data and model.

Many authors have provided comprehensive discussions of DIF detection methods (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Holland & Wainer, 1993; Millsap & Everson, 1993; Potenza & Dorans, 1995). The commonly-used nonparametric DIF detection methods for data containing a mixture of polytomous and dichotomous items are Mantel-Haenszel (MH, Mantel, 1963; Zwick, Donoghue, & Grima, 1993), Generalized Mantel-Haenszel (GMH), Standardized Mean Difference (SMD, Dorans & Schmitt, 1991) and Poly-SIBTEST (Chang et al. 1996). Commonly used parametric methods for mixed format data include Logistic Regression (LR, Swaminathan & Rogers, 1990), Ordinal Logistic Regression (OLR, Zumbo, 1999) and some Item Response Theory (IRT) based methods such as the Graded Response Model (GRM, Samejima, 1969) and the Generalized Partial Credit Model (GPCM, Muraki, 1992).

It has been shown that different DIF detection methods perform best under different conditions with the result that identification of DIF can vary according to which detection method is used. Factors that affect DIF detection rates include test length, sample sizes, ratio of the two groups’ sample sizes, amount of DIF, balance of DIF between the two groups, differences in item difficulty and discrimination, and differences between the ability levels of the two groups. For this reason, it has been suggested that DIF analyses should be conducted using two different DIF detection methods, and that only those items identified by both
methods should be flagged for potential DIF (Ercikan et al. 2004; Kim & Cohen, 1995). The two methods chosen to examine item level equivalence in this study are OLR and Poly-SIBTEST.

*Ordinal Logistic Regression*

As with many other DIF detection methods, logistic regression was originally developed for analyzing DIF in items that were scored dichotomously. Extensions to the original logistic regression methods have been made to accommodate polytomous items with ordinal data. For example, Miller and Spray (1993) described a form of logistic discriminant function analysis (LDFA) for DIF detection which was able to analyze both dichotomously and polytomously scored items simultaneously. Their method had some advantages over some of the previous DIF detection methods, in that both item scoring formats could be accommodated simultaneously, and both uniform and non-uniform DIF could be assessed. Although there was an associated chi-squared test of significance for the logistic discriminant function analysis method, there was no effect size measure.

Zumbo (1999) introduced an ordinal logistic regression method for analyzing DIF in items which are scored using an ordinal scoring format. He also introduced an effect size measure. His procedure, OLR, is an observed score parametric procedure which provides a hierarchical test of DIF that is able to distinguish uniform and nonuniform DIF. Zumbo suggests using a linear regression equation that regresses the variables of total score, group membership, and the interaction between total and group, (the predictor variables) on the dependent variable of score on an item, a latent, continuously-distributed random
variable. The regression is carried out in a hierarchical, or step-wise, fashion and a chi-squared value is calculated for each step of the regression as a test of significance. In addition, a coefficient of determination ($R^2$) is calculated at each step. Differences between $R^2$ values with the addition of each new step in the model give an indication of the amount of variance accounted for by the variable being added at that step, and this provides a step-wise effect size in addition to the test of significance.

Zumbo's (1999) OLR method was selected for use in this study because it has advantages over other DIF detection methods such as generalized Mantel-Haenszel and logistic discriminant function analysis (Zumbo, 2008). It can model both binary and ordinal items using the same modelling strategy, is relatively easy to implement and has both a test statistic and a measure of effect size. The OLR effect size, $R^2\Delta$, can be compared to the SIBTEST effect size, $\beta_u$. Jodoin and Gierl (2001) conducted a cubic curve regression to predict the OLR effect size from the previously established SIBTEST effect size measure thus establishing comparable effect sizes between the two methods.

Few empirical studies exist that draw specific comparisons between DIF detection results using OLR and Poly-SIBTEST. This research study will contribute to the research base by providing such a comparison.

*Poly-SIBTEST*

Shealy and Stout (1993) presented a procedure for multidimensional modeling of DIF (MMD) called SIBTEST, where SIB stands for "simultaneous item bias". Poly-SIBTEST is a modification of the SIBTEST procedure to
accommodate polytomous items. Both are nonparametric procedures that estimate the amount of DIF in an item and provide a statistical test of significance as well as an effect size measure. SIBTEST was designed to detect differential functioning in dichotomous items, and Poly-SIBTEST was designed to extend the detection method to include either dichotomous or polytomous items (Chang et al., 1996). Shealy and Stout's MMD model sets out that DIF is a result of a secondary ability or construct, called eta (η), being measured in addition to the ability of interest or main construct, theta (θ). In other words, there is more than one dimension (construct) being measured by the item and the studied groups differ in ability on the second dimension. Thus, when two groups who are matched on θ have different conditional ability distributions on η, DIF is indicated. In cases where there is a secondary dimension being measured on which there is no group difference in ability, DIF is not indicated. The presence of true ability distribution differences on the primary construct can statistically inflate results, and SIBTEST uses a regression correction to eliminate target ability difference from its test statistic.

A considerable amount of empirical research has shown that detection of DIF is comparable to other DIF detection methods such as Mantel-Haenszel and Logistic Regression (Jiang & Stout, 1998; Roussos & Stout, 1996; Shealy & Stout, 1993). In cases where there is a large difference in group ability levels, SIBTEST has been shown to be more accurate and powerful than other methods (for example, Narayanan & Swaminathan, 1994). Gierl, Jodoin, and Ackerman (2000) studied MH, SIBTEST, and LR under conditions in which there was a
large percentage of DIF items, and found that all three provide adequate protection against Type I error (false identification of items as being DIF when they are not) even when 60% of the items contain DIF, but SIBTEST was the most powerful. Considering the amount of research that indicates a large percentage of items display DIF in translated assessments, this result provides evidence that SIBTEST is an appropriate choice in these circumstances.

For this study which used SIBTEST’s polytomous extension, Poly-SIBTEST, it is appropriate to consider how Poly-SIBTEST compares to other DIF detection methods. Two modifications were made to SIBTEST to extend it to the polytomous case (Chang et al., 1996). One modification involved the inclusion of a multiplicative factor in the SIBTEST test statistic so that there would be more matching scores used in Poly-SIBTEST. The other modification was the substitution of coefficient α for the KR-20 which SIBTEST uses in its regression correction. The two methods are very closely related and there are many examples in the literature where they are treated as one method. However, they are technically different; therefore a separate literature review was conducted to determine the suitability of Poly-SIBTEST for this study.

Chang et al. (1996) introduced Poly-SIBTEST as an extension of the SIBTEST DIF detection method to include polytomous items. In their simulation study, Chang et al. compared Poly-SIBTEST to the Mantel Haenszel (MH) and Standardized Mean Difference (SMD) procedures. Their research showed that the Type 1 error rate of Poly-SIBTEST remained consistent when sample sizes were increased whereas MH and SMD did not. In addition, Poly-SIBTEST performed
better than MH and SMD when item discrimination parameters were set at
realistic values such as those from an actual administration of the National
Assessment of Educational Progress (NAEP) assessment.

In a simulation study comparing two parametric DIF procedures which use
the Graded Response Model (GRM) (Samejima, 1969) with the nonparametric
Poly-SIBTEST, Bolt (2002) established that Poly-SIBTEST resulted in greater
Type 1 error control when the data did not fit the parametric models used. Poly-
SIBTEST Type 1 error results were consistent across differences in sample size,
mean ability, and generating models. Poly-SIBTEST power rates were noticeably
lower than the two parametric procedures with low sample sizes (300 examinees)
but were slightly improved with sample sizes of 1000. Bolt concluded that Poly-
SIBTEST might offer advantages when the parametric DIF detection methods
cannot be used due to model misspecification.

A complex simulation study of polytomous DIF and violations of ordering
of the expected latent (OEL) trait by the raw score was conducted by DeMars
(2008). Although it might be expected that examinees ordered by observed score
would be similarly ordered on the expected value of the latent trait, this is not
always the case. DeMars varied the amount of violation of OEL, test length (5 or
20 items), differences in group abilities, and sample size (250 or 1000) and
compared Type 1 error results using four methods: Mantel, GMH, LDFA and
Poly-SIBTEST. Results showed that although Poly-SIBTEST displayed greater
error than the other procedures for the 5 item tests, it performed at least as well as
the other procedures for the 20 item tests. When the sample size was 1,000 per
group and the group means differed, Poly-SIBTEST had lower rates of Type I errors than the other procedures.

Zwick, Thayer, and Mazzeo (1997) conducted an analysis of Type 1 error control of the MH, SMD, and Poly-SIBTEST DIF detection methods. Their results showed that when reference and focal groups had identical ability distributions, SMD outperformed the other two methods. However, when the group abilities differed, Poly-SIBTEST produced the best results; and when the group abilities were different and item discrimination of the studied item was higher than that of the matching items, Poly-SIBTEST had much better Type 1 error control than the other methods.

Taken together, these studies of Poly-SIBTEST provide evidence for the appropriateness of its use in the current study. Poly-SIBTEST has been shown to exhibit good Type 1 error control when compared to both parametric and nonparametric methods. It performs at least as well as or better than other methods at sample sizes that approximate those used in this study, and in situations where group differences in ability might be expected, as in this study. In addition, it has both a test of significance and an effect size measure thus it lends itself to comparison with OLR on both of these criteria.

Establishing Bundle Level Equivalence: Differential Bundle Functioning Analyses

Statistical methods for analyzing bundles of items are more powerful than methods for analyzing single items (Gierl et al., 2001). Douglas, Roussos, and Stout (1996) proposed two methods of selecting bundles of items for DBF
analysis. They described a bundle as a cluster of items chosen according to some organizing principle. Based on Shealy and Stout’s (1993) multidimensional modeling of DIF (MMD), Douglas et al. proposed the use of expert opinion alone which provides substantive evidence in a confirmatory approach, or the use of statistical dimensionality analysis sometimes combined with expert opinion in an exploratory approach, in order to organize test items into bundles for differential functioning analysis. Roussos and Stout (1996) suggested a paradigm to unify the substantive and statistical DIF analyses by linking them to a theoretically sound and mathematically rigorous multidimensional conceptualization of DIF. When DIF hypothesis are generated based on substantive characteristics, or organizing principles, of the items (such as item content, differential task demands, or predicted group differences), which are then linked to the underlying structure of the data, bias can be minimized.

Two recent studies have been conducted to assess the performance and suitability of SIBTEST specifically for the purpose of analyzing DBF. Russell (2005) conducted simulation study analyses of two IRT-based techniques, Differential Fit of Items and Tests (“DFIT”, Raju, 1999 as cited in Russell, 2005) and SIBTEST in detecting differential test and differential bundle functioning respectively. His results, with regard to SIBTEST and DBF, suggest that SIBTEST performed with adequate power in his DBF power study. In contrast to other studies though, Russell determined that SIBTEST performed poorly when there was a large difference in ability between the target and reference groups.
This may be due to the fact that he evaluated SIBTEST performance at the conservative alpha level of 0.01 whereas others have used the 0.05 criteria.

Ross (2007) also assessed the performance of SIBTEST under various conditions in terms of its Type I error and power for DBF analysis with simulated data. Her study provides evidence that SIBTEST is able to distinguish between multidimensionality and multidimensionality that leads to DIF or DBF. In addition, Ross determined that SIBTEST has good power for detecting DBF, and controlled its Type I error rate for almost all of the simulated conditions. Her results regarding Type 1 error rate appear to conflict with those of Russell (2005) since Ross had elected to use a less stringent alpha level of 0.05.

Originating with the work of Nandakumar (1993) who showed that amplification, partial cancellation, or complete cancellation of DIF can occur when items are analyzed in groups or “bundles”, several studies have since demonstrated the use of SIBTEST’s bundle capabilities. Both confirmatory and exploratory analyses have been used to develop theory as to the potential causes for differential functioning in an assessment. Examples can be found in the work of Abbott (2007), Mendes-Barnett and Ercikan (2006), and Stout et al (2003). Fewer studies (Fouad & Walker, 2004; Gierl, 2005) have used Poly-SIBTEST for developing theory regarding causes of differential functioning. For example, Gierl (2005) demonstrated how Poly-SIBTEST can be used to identify and interpret sources of DIF by first conducting a substantive analysis of test items for organizing principles. This analysis can take the form of test specifications analysis, content analysis (based on expert opinion or literature review),
psychological analysis (such as cognitive tasks required) or empirical analysis of dimensionality. Based on the results of the substantive analysis, one can formulate a DIF or DBF hypothesis, that is, whether a bundle of items which are designed to measure a primary dimension also measure a secondary dimension. A statistical analysis can then be carried out to determine whether the organizing principles reveal distinct primary and secondary dimensions which lead to differential functioning. The current study provides additional research on the use of Poly-SIBTEST for developing differential functioning theory.

**Establishing Test Level Equivalence: Comparison of TCCs**

As with DIF studies, a variety of methods exist for the analysis of differential test functioning (DTF). These methods have been categorized by Pae and Park (2006) as methods that study the association between test score and an external criterion, IRT Test Characteristic Curve difference methods, and methods that compare internal factor structures across groups. This study examined item response theory test characteristic curves to investigate the degree of test level equivalence.

Item response theory methods can be used for the analysis of DIF as well as DTF. Item response theory provides us with mathematical models in which the probability of success on (or endorsing) an item varies with an examinee’s latent ability, theta (θ), and certain parameters that distinguish the item. According to IRT theory, the probability of success on a given item is a monotonically increasing function of examinees’ ability, which can easily be represented graphically for each item on a test, with θ plotted on the abscissa and the
probability of correct response on the ordinate axis. Item parameters determine the shape and positioning of the curve. This graphical representation is referred to as an Item Characteristic Curve (ICC). The sum of all the ICCs in a test give a Test Characteristic Curve (TCC).

There are three commonly-used item response models which are related to the number of item parameters being estimated. In the one-parameter logistic model (1PL), a single item parameter, item difficulty \( b \), is estimated, where difficulty is defined as the point on the ability scale where the probability of a correct response is 0.5. The two-parameter logistic model (2PL) provides an estimation of the item discrimination parameter \( a \) in addition to item difficulty. The \( a \) parameter is proportional to the slope of the ICC at the difficulty point, \( b \), on the ability scale. The higher the value of the discrimination parameter is, the steeper the slope of the ICC will be. The three-parameter logistic model (3PL) adds a parameter for “guessing” or “pseudo-chance” \( c \) parameter) which represents the probability of examinees with low ability correctly answering an item. The \( c \) parameter is seen as the lower asymptote on the item characteristic curve.

It is possible to investigate DTF directly by making comparisons of the item parameters for two or more groups, or indirectly by making comparisons of the TCCs for the groups. If DTF is not present, the TCCs generated separately from data from a focal group and a reference group will be identical. If the TCCs for the two groups are different, this is a representation of DTF between the groups.
Summary

Validity and score comparability are related to the uses and inferences drawn from assessment results. This chapter highlighted some of the uses and inferences from large-scale assessments in the Canadian context, with a particular focus on the PIRLS assessment, the 2001 administration of which is the measure that was analyzed in this study.

The responsibilities of test developers to provide empirical evidence for comparability of large-scale assessments which have been translated have been outlined, with reference to the Standards (AERA, APA & NCME, 1999) and the ITC Guidelines (ITC, 2001).

This chapter also provided a review of statistical methods that are appropriate for use in establishing score comparability at the item, bundle and test levels. The methods selected for use in this study were introduced. Their suitability for inclusion in this study was substantiated with thorough review and consideration of previous research.
CHAPTER III: METHOD

Overview

The purpose of this research study was to examine the degree of
differential item, bundle, and test functioning in the PIRLS 2001 Reader Booklet
and to draw comparisons across construct comparability analysis results at these
three levels. In addition, this study examined the similarities and differences
between two DIF detection methods: OLR and Poly-SIBTEST, and looked for
patterns or relationships between differential functioning analysis results at all
three levels of analysis.

This chapter presents a description of the PIRLS 2001 measure, including
the target population, construct intended to be measured, PIRLS assessment
framework, matrix sampling methodology and the manner in which items were
grouped by PIRLS into blocks and distributed in booklets. This chapter also
provides information regarding the data used in this study, specifically the Reader
booklet data for the four countries included in the study, sample sizes and
information regarding item types and their placement in two blocks in the Reader
booklet. Finally, descriptions of the analyses conducted at the item level (OLR
and Poly-SIBTEST), bundle level (Poly-SIBTEST) and scale level (comparison of
TCCs) are given.

Measure

The PIRLS 2001 assessment was developed and conducted by the
International Association for the Evaluation of Educational Achievement (IEA). It
was developed to provide policymakers, educators, researchers and practitioners
with information about how well students are mastering reading literacy, and to provide information for improvement. Thirty-five countries joined together to conduct the first PIRLS assessment in 2001. The assessment was originally constructed in English and was then translated/adapted and administered in 28 different languages (Martin, Mullis, & Kennedy, 2003).

The international target population for PIRLS 2001 was “all students enrolled in the upper of the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing” (Foy & Joncas, 2003, p. 54). In most, but not all countries, this represented the fourth grade. PIRLS 2001 chose this population because the age of 9 years represents an important transition point in the development of reading skills where students are expected to have learned how to read and are now reading to learn. In addition, the PIRLS 2001 study was intended to complement another study, Trends in International Mathematics and Science Study (TIMSS) which assesses math and science achievement at the same grade (Grade 4), thus providing countries who take part in both assessments with data in three subjects at the Grade 4 level.

PIRLS 2001 was designed to assess the construct of “reading literacy” which is defined as

the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers, and for enjoyment. (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001, p. 3)
This definition reinforces the concept that reading is a complex and interactive process. The PIRLS assessment framework provides a theoretical basis for understanding reading literacy, whose definition was derived from and informed by numerous theories of reading (Sainsbury & Campbell, 2003, p 15). There are three specific aspects of reading literacy which are the focus of PIRLS 2001. They are processes for comprehension, purposes for reading, and reading behaviors and attitudes. Reading behaviors and attitudes were assessed through the use of four questionnaires which were developed to gather contextual information on home and school experiences associated with developing reading literacy, as well as home and school environmental, demographic and socioeconomic factors. Students, their parents, teachers and school principals were surveyed.

Table 1 presents the PIRLS 2001 aspects of reading which form the foundation for the PIRLS written achievement assessment, the focus of this study. They are processes for comprehension and purposes for reading.
Each PIRLS 2001 item was written to engage students in one of the two purposes for reading and one of the four processes for comprehension (Sainsbury & Campbell, 2003, p 17). The four processes for comprehension are composed of a hierarchical sequence from simple retrieval, to making inferences, interpreting and integrating information, and finally examining and evaluating elements of what is being read. These four processes require increasing meta-cognitive processes and strategies that allow readers to examine their understanding and adjust their approach. In addition, the knowledge and experience that readers bring to reading equip them with an understanding of language, texts, and the world through which they filter their comprehension of the material. In order to respond to questions intended to assess the two higher cognitive levels (interpreting and

<table>
<thead>
<tr>
<th>Processes for Comprehension</th>
<th>Focus on and retrieve explicitly stated information.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Make straightforward inferences.</td>
</tr>
<tr>
<td></td>
<td>Interpret and integrate ideas and information.</td>
</tr>
<tr>
<td></td>
<td>Examine and evaluating content, language, and textual elements.</td>
</tr>
<tr>
<td>Purposes for Reading</td>
<td>Reading for literary experience</td>
</tr>
<tr>
<td></td>
<td>Reading to acquire and use information</td>
</tr>
</tbody>
</table>

Table 1

*PIRLS 2001 Aspects of Reading*
integrating information, and examining and evaluating elements), readers may need to depend on background knowledge and personal experience (Campbell et al. 2001). Overall, the PIRLS assessment contained a balanced representation of items in each of the four process categories, with each one representing either 20 or 30% of the total items. The two purposes of reading focused on in the written assessment are reading for literary experience and reading to acquire and use information.

The PIRLS 2001 reading literacy assessment consists of selected reading passages of approximately 1000 words, with associated questions in two question formats, multiple choice and constructed-response. Reading passages are either literary or informational in nature. In total, the assessment contains 8 blocks of items (4 literary and 4 informational blocks) for a total of 98 items, with a score value of 133 points, distributed across 10 booklets which were matrix sampled. Each student was assigned only 1 of the 10 booklets.

When the purpose of testing is to measure the performance of groups rather than individuals, matrix sampling procedures are frequently used. In matrix sampling, a small subset of items is assigned to each of many subsamples of examinees. Data are aggregated across subsamples and item subsets to obtain a measure of group performance (AERA, APA & NCME, 1999). Matrix sampling allows for broader coverage of the construct while at the same time reducing the amount of time required for each respondent to take the assessment. Student performance on the entire set of items is then inferred based on completing just a subset of the items.
Three literary and three informational blocks of items were linked among booklets 1 to 9 through a sophisticated linking procedure. For example, items that comprised block "L3" were responded to by students who received Booklets 2, 3 and 9. The fourth literary and informational blocks were contained only in the Reader booklet. The distribution of blocks is shown in Table 2:

Table 2

*PIRLS 2001* 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>6</td>
<td>L3</td>
</tr>
<tr>
<td>2</td>
<td>L1</td>
<td>7</td>
<td>L2</td>
</tr>
<tr>
<td>3</td>
<td>L2</td>
<td>8</td>
<td>L1</td>
</tr>
<tr>
<td>4</td>
<td>L3</td>
<td>9</td>
<td>L2</td>
</tr>
<tr>
<td>5</td>
<td>L1</td>
<td>Reader</td>
<td>L4</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td></td>
<td>I4</td>
</tr>
</tbody>
</table>

Booklet 10, called the "PIRLS 2001 Reader Booklet", was unique in many respects. Its uniqueness led to it being chosen as the source of data for the current study, as outlined next.

The Reader booklet contained 2 blocks (1 literary and 1 informational) of items that were not linked to any other blocks or booklets. Rather, the Reader items were linked to the rest of the assessment through the use of randomly equivalent samples. Booklets 1 to 9 were printed in black and white and had reading passages and questions contained within a single leaflet. In contrast, the
Reader booklet was printed in full colour in an attractive magazine-style format. The Reader was distributed at three times the rate of the other 9 booklets thus giving it a three-fold increase in sample size over any other individual booklet. Its reading passages and questions were contained in two separate booklets. The Reader was thought to be a natural, authentic reading setting for 9-year-olds. A natural implication of its intended authenticity is that the Reader may be a key booklet for capturing the construct of reading literacy. A last distinction of the Reader booklet that makes it an appropriate choice for study is that both of its reading passages and all associated questions have been released into the public domain and made available to the public in English (Gonzalez & Kennedy, 2003). The Reader Booklet passages and questions are contained in Appendix A attached to this document. The time allotted to complete the reading literacy assessment was 80 minutes, and a further 20 minutes was allotted for each student to complete the student questionnaire.

PIRLS 2001 written reading literacy achievement results are reported in the PIRLS 2001 International Report (Mullis, Martin, Gonzalez, & Kennedy, 2003). This report contains assessment results by country for each of the two purposes of reading, and for overall reading achievement. The PIRLS International Achievement in the Processes of Reading Comprehension report (Mullis, Martin, & Gonzalez, 2004) provides achievement results for the comprehension processes. Comparisons are made between countries, and information is given as to whether each country is significantly higher or lower than the international average. Achievement results by gender were also reported,
as were the percentages of students in each country reaching defined international benchmarks in reading achievement. The sample size, mean proficiency on overall reading achievement, years of formal schooling, and average age for each of the participating countries is shown in Table 3.
<table>
<thead>
<tr>
<th>Country</th>
<th>Sample Size</th>
<th>Overall Average Scale Score</th>
<th>Years of Formal Schooling</th>
<th>Average Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>6044</td>
<td>561 (2.2)</td>
<td>4</td>
<td>10.8</td>
</tr>
<tr>
<td>Netherlands</td>
<td>4112</td>
<td>554 (2.5)</td>
<td>4</td>
<td>10.3</td>
</tr>
<tr>
<td>England</td>
<td>3156</td>
<td>553 (3.4)</td>
<td>5</td>
<td>10.2</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>3460</td>
<td>550 (3.8)</td>
<td>4</td>
<td>10.9</td>
</tr>
<tr>
<td>Latvia</td>
<td>3019</td>
<td>545 (2.3)</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Canada (O,Q)</td>
<td>8253</td>
<td>544 (2.4)</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Lithuania</td>
<td>2567</td>
<td>543 (2.6)</td>
<td>4</td>
<td>10.9</td>
</tr>
<tr>
<td>Hungary</td>
<td>4666</td>
<td>543 (2.2)</td>
<td>4</td>
<td>10.7</td>
</tr>
<tr>
<td>United States</td>
<td>3763</td>
<td>542 (3.8)</td>
<td>4</td>
<td>10.2</td>
</tr>
<tr>
<td>Italy</td>
<td>3502</td>
<td>541 (2.4)</td>
<td>4</td>
<td>9.8</td>
</tr>
<tr>
<td>Germany</td>
<td>7726</td>
<td>539 (1.9)</td>
<td>4</td>
<td>10.5</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>3022</td>
<td>537 (2.3)</td>
<td>4</td>
<td>10.5</td>
</tr>
<tr>
<td>New Zealand</td>
<td>2488</td>
<td>529 (3.6)</td>
<td>5</td>
<td>10.1</td>
</tr>
<tr>
<td>Scotland</td>
<td>2717</td>
<td>528 (3.6)</td>
<td>5</td>
<td>9.8</td>
</tr>
<tr>
<td>Singapore</td>
<td>7002</td>
<td>528 (5.2)</td>
<td>4</td>
<td>10.1</td>
</tr>
<tr>
<td>Russian Fed.</td>
<td>4093</td>
<td>528 (4.4)</td>
<td>3 or 4</td>
<td>10.3</td>
</tr>
</tbody>
</table>
Table 3 (Continued)

<table>
<thead>
<tr>
<th>Country</th>
<th>Sample Size</th>
<th>Overall Average Scale Score (Standard Error)</th>
<th>Years of Formal Schooling</th>
<th>Average Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hong Kong, SAR</td>
<td>5050</td>
<td>528 (3.1)</td>
<td>4</td>
<td>10.2</td>
</tr>
<tr>
<td>Greece</td>
<td>2494</td>
<td>524 (3.5)</td>
<td>4</td>
<td>9.9</td>
</tr>
<tr>
<td>Slovak Republic</td>
<td>3807</td>
<td>518 (2.8)</td>
<td>4</td>
<td>10.3</td>
</tr>
<tr>
<td>Iceland</td>
<td>3676</td>
<td>512 (1.2)</td>
<td>4</td>
<td>9.7</td>
</tr>
<tr>
<td>Romania</td>
<td>3625</td>
<td>512 (4.6)</td>
<td>4</td>
<td>11.1</td>
</tr>
<tr>
<td>Israel</td>
<td>3973</td>
<td>509 (2.8)</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Slovenia</td>
<td>2952</td>
<td>502 (2.0)</td>
<td>3</td>
<td>9.8</td>
</tr>
<tr>
<td><strong>Intern'l Avg.</strong></td>
<td><strong>500 (0.6)</strong></td>
<td></td>
<td><strong>4</strong></td>
<td><strong>10.3</strong></td>
</tr>
<tr>
<td>Norway</td>
<td>3459</td>
<td>499 (2.9)</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Cyprus</td>
<td>3001</td>
<td>494 (3.0)</td>
<td>4</td>
<td>9.7</td>
</tr>
<tr>
<td>Moldova, Rep.</td>
<td>3533</td>
<td>492 (4.0)</td>
<td>4</td>
<td>10.8</td>
</tr>
<tr>
<td>Turkey</td>
<td>5125</td>
<td>449 (3.5)</td>
<td>4</td>
<td>10.2</td>
</tr>
<tr>
<td>Macedonia, Rep.</td>
<td>3711</td>
<td>442 (4.6)</td>
<td>4</td>
<td>10.7</td>
</tr>
<tr>
<td>Colombia</td>
<td>5131</td>
<td>422 (4.4)</td>
<td>4</td>
<td>10.5</td>
</tr>
<tr>
<td>Argentina</td>
<td>3300</td>
<td>420 (5.9)</td>
<td>4</td>
<td>10.2</td>
</tr>
<tr>
<td>Iran, Islamic Rep.</td>
<td>7430</td>
<td>414 (4.2)</td>
<td>4</td>
<td>10.4</td>
</tr>
<tr>
<td>Kuwait</td>
<td>7133</td>
<td>396 (4.3)</td>
<td>4</td>
<td>9.9</td>
</tr>
<tr>
<td>Morocco</td>
<td>3153</td>
<td>350 (9.6)</td>
<td>4</td>
<td>11.2</td>
</tr>
<tr>
<td>Belize</td>
<td>2909</td>
<td>327 (4.7)</td>
<td>4</td>
<td>9.8</td>
</tr>
</tbody>
</table>

*Canada is represented by the provinces of Ontario and Quebec only. The international average does not include the results from these provinces separately.*
Data

This study analyzed PIRLS 2001 Reader booklet data from four of the thirty-five participating countries: Argentina, Colombia, England and the United States. These countries were chosen based on their relative positions in the distribution of achievement results from PIRLS 2001, the language in which they administered the assessment, and potential cultural similarities and differences. Argentina and Colombia both ranked below the international average of all participating countries (fifth and sixth-lowest respectively), while England and the United States ranked well above the international average (third and ninth-highest respectively). Argentina and Colombia were the only two countries that translated the assessment into Spanish and administered it in Spanish, while England and the United States administered it in English, requiring no translation. This selection of countries to analyze was made based on the fact that, if differential functioning was present (whether at the item, bundle, or test level), it would be most likely to be found between high and low ranking countries. In addition, a significant amount of research has indicated that differential functioning may be a result of many factors such as translation, culture, and socio-economic status. Each of these factors may therefore contribute to the presence of differential functioning in this study.

As shown in Table 3, PIRLS 2001 was administered to 146,490 students, whose average age was 10.3 years, and who had had an average of 4 years of formal schooling. The sample sizes for the four countries analyzed in this study are also shown, and range from 3,156 to 5,131. Students in England had received,
on average, 5 years of formal schooling compared with 4 years in Argentina, Colombia and the United States. Students in Colombia were an average age of 10.5 years, compared with 10.2 years in the other three countries.

Since the PIRLS 2001 booklets were matrix sampled, not all students took the Reader Booklet and its associated items. The average sample size per country for the Reader Booklet was 1000 students. Table 4 shows the sample sizes for the PIRLS 2001 Reader Booklet only for the four countries analyzed in this study.

Table 4

<table>
<thead>
<tr>
<th>Country</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sample</td>
<td>35 987</td>
</tr>
<tr>
<td>Argentina</td>
<td>794</td>
</tr>
<tr>
<td>Colombia</td>
<td>1 285</td>
</tr>
<tr>
<td>England</td>
<td>777</td>
</tr>
<tr>
<td>USA</td>
<td>926</td>
</tr>
</tbody>
</table>

This study investigated score comparability for four countries: Argentina, Colombia, England, and USA. Data from each of these countries were analyzed separately as a focal group, which were compared to a reference group. The reference group was composed of a random sample of 200 students from each of the 34 remaining countries. For example, the analyses for England compared the entire sample of English students who wrote the Reader Booklet (the focal group) against an “International Sample” which consisted of a random sampling of
students who wrote the Reader booklet from all countries except for England (the reference group). This resulted in a reference group size of 6800. The reference group sample creation followed the method used by PIRLS 2001 in creating an international sample for scaling as closely as possible. The international sample is intended to be representative of all individual countries.

As with all other PIRLS 2001 booklets, the Reader booklet contains two reading passages with their associated questions. One reading passage, “Hare Heralds the Earthquake”, is intended to assess reading for literary experience, and the other, “Nights of the Pufflings”, is intended to assess reading to acquire and use information. In total, there were 24 items in the Reader associated with these two passages, 13 of which were multiple choice and 11 of which were constructed response items, as shown in Table 5. The corresponding item numbering used throughout this study is also shown in Table 5, in the column entitled “Study Item #.”
Table 5

**PIRLS 2001 Reader Booklet Items**

<table>
<thead>
<tr>
<th>Reading Passage</th>
<th>PIRLS Item #</th>
<th>Study Item #</th>
<th>Item Type*</th>
<th>Max Score (points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hare Heralds the</td>
<td>1</td>
<td>1</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td>Earthquake</td>
<td>2</td>
<td>2</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>CR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td>CR</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>7</td>
<td>CR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>8</td>
<td>CR</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9</td>
<td>CR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>CR</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>11</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td>Nights of the</td>
<td>1</td>
<td>12</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td>Pufflings</td>
<td>2</td>
<td>13</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>14</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>15</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>16</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>17</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>18</td>
<td>CR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>19</td>
<td>CR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>20</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>21</td>
<td>CR</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>22</td>
<td>MC</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>23</td>
<td>CR</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>24</td>
<td>CR</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total Score:** 32 Points

*MC = Multiple Choice; CR = Constructed Response*
Analyses

Several statistical analyses were conducted to investigate score comparability in the PIRLS 2001 Reader booklet. Analyses were conducted for comparability at the item, bundle and scale levels. These analyses are described below.

Item Level Analyses

Comparability at the item level was investigated using both Ordinal Logistic Regression and Poly-Sibtest analyses.

Ordinal Logistic Regression

Zumbo’s (1999) OLR method uses a linear regression equation which regresses the variables of total score, group membership, and the interaction between total and group (the predictor variables) on the dependent variable of score on an item, a latent, continuously-distributed random variable. The OLR equation is

\[ y^* = b_0 + b_1 \text{TOTAL} + b_2 \text{GRP} + b_3 \text{TOTAL} \times \text{GRP} + \epsilon_i \]

where \( y^* \) is the continuously-distributed random variable being predicted (in this study, \( y^* \) represents the predicted score on an item), \( b_0 \) represents the slope of the regression line, and \( b_1, b_2, \) and \( b_3 \) are the regression coefficients (slopes, or beta coefficients) of each of the predictor variables total score, group membership, and total score by group membership interaction. The error terms, \( \epsilon_i \), are distributed with a mean of zero and variance of \( \pi^2/3 \).

Zumbo (1999) sets out that OLR assumes an unobservable continuum of variation and that it can be expressed as:
\[
\log \left[ \frac{P(Y \leq j)}{P(Y > j)} \right] = \alpha_j + b(X) \quad \text{or}
\]

\[
\logit[P(Y \leq j)] = \alpha_j + b(X)
\]

where a logit is the natural logarithm of the ratio of two probabilities and \( j = 1, 2, \ldots, c-1 \), where \( c \) is the number of categories in the ordinal scale. The model requires a separate intercept parameter \( \alpha_j \) for each cumulative probability. For a 3-point response, an OLR model describes the effect of \( X \) (the total score for the scale) on the odds that \( Y \leq 1 \) instead of \( Y > 1 \) on the scale, and the effect of \( X \) on the odds that \( Y \leq 2 \) instead of \( Y > 2 \). For DIF analysis, the OLR equation becomes:

\[
\logit[P(Y \leq j)] = \alpha_j + b1_{\text{TOTAL}} + b2_{\text{GRP}} + b3_{\text{TOTAL} \times \text{GRP}}
\]

The selection of a matching variable is an important decision to be made when implementing an OLR analysis because logistic regression models the probability of responding correctly to an item by group membership and a matching variable, often the total scale score for the assessment. The matching variable used in this study was a "rest score" which is the total score excluding the item under investigation. Slocum, Gelin, and Zumbo (in press) suggest that for short scales a rest score matching criteria be used. The PIRLS 2001 Reader booklet consists of 24 items only which is considered to be a short scale.

Ordinal logistic regression is carried out in a hierarchical, or step-wise, fashion with the order of entry of variables determined by the researcher, first entering total score. The second step involves holding the effect of the
conditioning variable (the total score) constant, while measuring the effect of the
group membership variable. This step is representative of uniform DIF. In the
third step, the first two variables are held constant while the effect of the
interaction is measured. This step represents non-uniform DIF, that is DIF that
varies along the ability continuum. A chi-square value is calculated for each step
of the regression as a test of significance. For this study, a significance level of p
< .01 was set as the criterion for the chi-square test.

In addition, the coefficient of determination, Nagelkerke R², was
calculated at each step. Differences between R² values with the addition of each
new step in the model give an indication of the amount of variance accounted for
by the variable being added at that step, and this provides an effect size, R²Δ, in
addition to the chi-square test of significance. Jodoin and Gierl’s (2001) effect
size classification scheme will be used in the current study:

Negligible DIF: R²Δ < .035

Moderate DIF: Chi-squared null hypothesis is rejected
and .035 < R²Δ < .070

Large DIF: Chi-squared null hypothesis is rejected
and R²Δ > .070

These effect size criteria have been chosen for this study because they have
been shown to be comparable to the effect size criteria for SIBTEST (Jodoin &
Gierl, 2001).
Poly-SIBTEST

SIBTEST (Shealy & Stout, 1993) is a latent variable, nonparametric approach to DIF detection. That is, groups are matched on an estimate of a latent variable, \( \theta \), instead of on their observed scores, and no functional form for the item response function is assumed.

In order to analyze data in SIBTEST or Poly-SIBTEST, the items must be separated into two subtests. Items hypothesized to display DIF are placed into a "studied subtest". In the Multidimensional Model for DIF, these would be items predicted to contain the primary dimension and at least one secondary dimension. Items believed to be free of DIF, that is, measuring only a primary dimension, are placed into the "matching subtest" (also called a "valid subtest"). The matching subtest is used for the purposes of matching examinees on the latent variable.

For dichotomous items, an item does not exhibit DIF if regression scores for the two groups are expected to be the same for all values of \( \theta \):

\[
E_R[Y|\theta] = E_F[Y|\theta]
\]

Equation (1)

where \( R \) refers to the reference group and \( F \) refers to the focal group under study.

Chang et al. (1996) made two modifications to SIBTEST to make it applicable to polytomous items, as well as a mixture of dichotomous and polytomous items. One modification was to allow for more score categories, since polytomous items are scored in \( m+1 \) ordered categories. A polytomous item can be shown not to exhibit DIF if the probability of achieving a score of \( k \) for a
randomly sampled examinee (with the same proficiency on \( \theta \)) from each group is the same:

\[ P_{i\theta} (\theta) = P_{i\theta'} (\theta), k = 1, \ldots, m. \]  

Equation (2)

Here, \( m \) refers to the maximum possible score for the items. Equation (1) has been shown to imply Equation (2) for the commonly used polytomous item response models, thus Equation (1) can be adopted for the latent variable definition of null DIF for ordinally scored polytomous items (Chan et al.). Poly-SIBTEST measures the amount of DIF at \( \theta \) (known as local DIF) as:

\[ B_\theta = E_{R}[Y|\theta] - E_{F}[Y|\theta] \]

Poly-SIBTEST also provides a global index of DIF as:

\[ \beta = \int B_\theta (\theta) f_{Y_i}(\theta) d\theta \]

where \( f_{Y_i}(\theta) \) is the density of \( \theta \) in the focal group. In this equation \( \beta \), representing effect size, is interpreted as the expected amount of DIF that a randomly selected focal group examinee would experience.

It has been shown that for the partial credit model, the generalized partial credit model, and the graded response model, matching test true scores (from classical test theory) are a one-to-one transformation of the latent variable \( \theta \) (Chang, 1994 and Chang & Mazzeo, 1994, as cited in Chang et al. 1996). Local DIF can then be estimated by:

\[ d_k = \bar{Y}_{ik} - \bar{Y}_{i\theta k}, k = 0, \ldots, n_H \]
50

where $\bar{Y}_{jk} - \bar{Y}_{jk}$ is the group difference in performance on the studied item among examinees with the same observed score, and an estimation of DIF is given by

$$\hat{\beta} = \sum_{k=0}^{n_u} p_k d_k$$

where $p_k$ is the proportion of all examinees achieving a score of $k$ on the matching test.

Finally, the Poly-SIBTEST test statistic is defined by:

$$\beta = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}$$

where

$$\hat{\sigma}(\hat{\beta}) = \sqrt{\sum_{k=0}^{n_u} p_k^2 \left( \frac{\hat{\sigma}^2(Y|k, R)}{N_{jk}} + \frac{\hat{\sigma}^2(Y|k, F)}{N_{jk}} \right)}$$

with $\hat{\sigma}^2(Y|k, R)$ and $\hat{\sigma}^2(Y|k, F)$ being the sample variances of the studied item scores for examinees in the reference and focal groups respectively.

It is important to note that the above equations were all formulated with the assumption that the focus and reference groups have equal target ability distributions. This is, of course, not always the case and the presence of unequal target abilities would lead to impact which must be distinguished from bias.

Shealy and Stout (1993) therefore used a regression equation to correct for target ability distribution differences. The matching test true scores for the groups are estimated by using the equation for the linear regression of true score on observed score. In the dichotomous SIBTEST case, the slope of the regression line was
estimated through the use of KR-20; for Poly-SIBTEST, the slope is calculated using Cronbach's alpha (Chang et al. 1996).

Poly-SIBTEST is a statistical test for testing the null hypothesis (H₀: β = 0) against the alternative hypothesis of DIF/DBF against the focal group (H₁: β > 0) or against the reference group (H₁: β < 0) or against either group (H₁: β ≠ 0). This study tested the alternative hypothesis of DIF against either group. When β is statistically significant and has a positive value, this is representative of DIF against the focal group, whereas if β is significant and negative this indicates DIF against the reference group.

For this analysis at the item level, a "standard" one-item-at-a-time DIF analysis using Poly-SIBTEST was conducted as explained in the software manual (The Roussos-Stout Software Development Group, n.d., p. 13) by analyzing each item as the suspect subset and the rest of the test as the matching subset. For this study, a significance level of p < .01 was set in order to maintain comparability with the OLR method that is being used. In addition, the guessing parameter was set to 0 for consistency with the test level analyses which are being conducted as part of this study, since TCCs for the scale level analysis were generated using a 2PL model (no guessing) for the dichotomous items.

Roussos and Stout (1996) proposed a classification scheme for interpreting the value of β for an item. Their proposed effect size classification is:
Negligible DIF: $|\beta| < .059$

Moderate DIF: $0.059 < |\beta| < 0.088$

Large DIF: $|\beta| > 0.088$

This effect size classification scheme was adopted for this study.

The results of DIF analyses using both OLR and Poly-SIBTEST were examined to seek out organizing principles on which to formulate a differential bundle functioning hypothesis, as outlined in the next section.

**Bundle Level Analysis**

Poly-SIBTEST was used in this study for analyzing DBF as well as for DIF. This study provided additional research on the use of Poly-SIBTEST for developing differential functioning theory. The results of DIF analyses using both OLR and Poly-SIBTEST were examined. The processes for reading comprehension of the items jointly identified as DIF by the two methods were identified (using test specifications set out in Campbell et al., 2001) to seek out organizing principles on which to group items into bundles for DBF as outlined in Douglas et al. (1996) and Gierl (2005). Once organizing principles were established, the items were grouped into bundles and analyzed using Poly-SIBTEST.

Although classification guidelines exist for interpreting effect size for DIF using SIBTEST and Poly-SIBTEST, no such guidelines exist to date for the interpretation of effect size for bundles of items (Boughton, Gierl, & Khaliq, 2000;
In the absence of established effect size guidelines for bundles, Gierl and Bolt proposed guidelines specifically for bundles of 5 items or less, being $|\beta| < .10$ for negligible DBF and $|\beta| \geq .10$ for moderate to large DBF. These proposed guidelines were followed if bundles of 5 items or less were distinguished in the current research.

**Scale Level Analyses**

This study compared test characteristic curves (TCCs) based on each international reference sample (as defined earlier in this chapter) versus its respective country (focal) group data to investigate the degree of comparability at the scale level and to investigate the appropriateness of using the international item parameters for establishing country scores. Recall that TCCs are a graphical representation of the expected probability of test score plotted against ability, and are based on item parameters estimated using item response theory. Also recall from Chapter 1 that current scaling methods in large-scale assessments involve the estimation of international item parameters from a data set containing samples of data from all countries. The estimated international item parameters are then used to create individual country scores. An assumption is made that the international item parameters are representative of all countries and are suitable for creating country scores. Differences in TCCs between the international reference sample and the country data may signify inappropriateness of the international item parameters for estimating country scores (Ercikan & Gonzalez, 2008).

Item response theory parameters were estimated for the international sample and for each of the four countries separately using PARDUX (Burket,
1991). PARDUX generates parameters simultaneously for dichotomous and polytomous items using marginal maximum likelihood procedures implemented via an EM algorithm, and can be used to estimate parameters with a variety of models, including the 2PL and PPC models (Fitzpatrick et al. 1996) which were used in the current study.

The 2PL IRT model which is used to jointly estimate the difficulty and discrimination parameters of items and ability is given by the equation

\[ P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \]

where \( P_i(\theta) \) is the probability that a randomly chosen examinee with ability \( \theta \) answers item \( i \) correctly, \( b_i \) is the item difficulty parameter, and \( a_i \) is the item discrimination parameter (Hambleton et al, 1991). As explained previously, the probability of a correct response to an item is a monotonically-increasing, s-shaped curve, with values between 0 and 1 over the ability range, the ICC function. The sum of all ICCs over a test results in the TCC.

The two-parameter partial credit (2PPC) model is explained in Ercikan et al. (1998) and can be summarized as a special case of Bock's nominal model (Bock, 1972) which is equivalent to Muraki's (1992) generalized partial credit model. Bock's nominal model states that an examinee with \( \theta \) ability, obtaining a score at the \( k \)th level on the \( j \)th item is:

\[ P_j(\theta) = P(X_j = k \mid \theta) = \frac{\exp \frac{Z_{jk}}{m_j}}{\sum_{k=1}^{m_j} \exp \frac{Z_{jk}}{m_j}}, k = 1, \ldots, m_j \]
where,

\[ Z_{jk} = A_{jk} \theta + C_{jk}. \]

The 2PPC model is a special case in that it includes the following constraints:

\[ A_{jk} = \alpha_j (k - 1), \]

and,

\[ C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}. \]

In this model, \( \gamma_{j0} = 0 \), and \( \alpha_j \) and \( \gamma_{ji} \) are free parameters to be estimated by the data. The first constraint implies that items can vary in their discriminations and that higher item scores reflect higher ability levels. In the 2PPC model, for each item there are \( m_j - 1 \) independent \( \gamma_{ji} \) difficulty parameters and one \( \alpha_j \) discrimination parameter; a total of \( m_j \) independent item parameters are estimated (Ercikan et al., p. 142).

Item response theory item parameters based on separate country calibrations and the international calibration are not on the same scale due to the arbitrariness of the IRT scales. Before comparisons can be made, the scales need to be linked to be placed onto a common scale. In this study linking was created using the robust mean and sigma transformation described by Stocking and Lord (1983). The scale of reference was that from the calibration done using the combined data sets from all countries including all items in the transformation (Ercikan & Gonzalez, 2008).
TCCs were computed and their representations were created using an Excel-based software, MODFIT (Stark, 2001, 2002). Differences in the percent of maximum score estimated between individual countries and the international sample were determined and compared.

**Model-Data Fit**

A model-data fit using Yen's (1984) Q1 fit statistic was calculated using PARDUX (Burket, 1991). The Q1 is a chi-square statistic used to test the goodness of fit between an IRT model and test data. Q1 is affected by the number of score points for each item as well as by the sample size (Fitzpatrick, Lee, & Gao, 2001). Therefore, Q1 is transformed into a standardized z-statistic. It is computed based on the squared differences in probabilities of responding correctly to an item based on observed data versus predicted by the item parameters and added across deciles on the score scale. As described in Fitzpatrick et al., items are considered to fit poorly if |z| ≥ 4.0 for sample sizes of 1,500. For larger sample sizes the critical value for determining poor fit is adjusted by a factor of N/1500 where N is the calibration sample size. For the purposes of this analysis where the international sample size is N=6800, a criterion of |z| ≥ 18 was used.

**Factor Structures**

The Exploratory Factor Analysis using Principal Components Analysis (PCA) was conducted to examine the factor structures of the PIRLS 2001 Reader booklet data separately for the four countries. The similarity of factor structures provides evidence of similarity of constructs measured at the scale level in the four country administrations of the PIRLS Reader. The underlying structure and
number of factors in the data for each country were examined using the Kaiser-Guttman rule, the percentage of variance accounted for by the factors, and examination of scree plots.

Summary

In this chapter the PIRLS 2001 measure was described, with a particular focus on the Reader Booklet. The target population, sample sizes, item and booklet formats were explained. The construct intended to be assessed has been discussed in detail. The statistical analyses and procedures used in this study were outlined, including the use of OLR and Poly-SIBTEST for item level analyses, Poly-SIBTEST for bundle level analysis, and TCC comparisons for analysis at the scale level. The next chapter presents the results of these analyses with a view to answering the research questions that are the focus of this study.
CHAPTER IV: RESULTS

This chapter presents the results of the study, beginning with descriptive statistics of the PIRLS 2001 Reader booklet data. This is followed by results of the item level analyses using OLR and Poly-SIBTEST, together with comparison of the similarities and differences of the two DIF detection methods. The manner in which items were bundled using the results of the DIF analyses is explained, and the results from the bundle level analysis are then presented. Finally, the results of the scale level analysis of TCC comparisons to investigate DTF and the appropriateness of using the international item parameters for establishing country scores are presented.

Descriptive Statistics of the PIRLS 2001 Reader Booklet Data

The Reader Booklet was written by a total of 35,987 students from the 35 participating countries. This study analyzed data from all students who took the PIRLS 2001 Reader booklet in the countries of Argentina, Colombia, England and USA. In order to create reference group comparison data for differential functioning studies, a random sample of 200 students from each of the remaining 34 participating countries was drawn, resulting in 4 reference group samples, each composed of 6800 examinees. The sample sizes and descriptive statistics for reference and focal groups are shown in Table 6.

The Reader booklet contained 24 items (13 multiple choice and 11 constructed response items) associated with two reading passages. The maximum possible score for the booklet is 32 points. The mean raw scores for Argentina (12.02) and Colombia (12.36) are below their respective reference group mean.
raw scores, with standardized score differences of 0.78 and 0.74. The mean scores for England (20.73) and USA (20.26) are both above their reference group mean scores, with standardized score differences of -0.47 and -0.41 respectively.

Table 6

*PIRLS 2001 Reader Booklet Descriptive Statistics*

<table>
<thead>
<tr>
<th>Country</th>
<th>Group</th>
<th>Sample</th>
<th>Standardized Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Size</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Difference&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Argentina</td>
<td>Reference</td>
<td>6800</td>
<td>17.61 (7.17)</td>
</tr>
<tr>
<td></td>
<td>Focal</td>
<td>794</td>
<td>12.02 (6.73)</td>
</tr>
<tr>
<td>Colombia</td>
<td>Reference</td>
<td>6800</td>
<td>17.59 (7.21)</td>
</tr>
<tr>
<td></td>
<td>Focal</td>
<td>1285</td>
<td>12.36 (6.05)</td>
</tr>
<tr>
<td>England</td>
<td>Reference</td>
<td>6800</td>
<td>17.33 (7.22)</td>
</tr>
<tr>
<td></td>
<td>Focal</td>
<td>777</td>
<td>20.73 (6.79)</td>
</tr>
<tr>
<td>USA</td>
<td>Reference</td>
<td>6800</td>
<td>17.35 (7.22)</td>
</tr>
<tr>
<td></td>
<td>Focal</td>
<td>926</td>
<td>20.26 (6.50)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Reference group – Focal group

The score distributions for each of the four countries are shown in Figures 1 and 2. These figures, which show the percentage of students achieving each score level in the four countries, demonstrate similar score distribution patterns for Argentina and Colombia and similar score distribution patterns for England and USA. A majority of Argentinean and Colombian students score in the lower half of the score range, while the majority of English and American students score in the upper half of the range. In fact, 72.8% and 74.2% of students from Colombia
and Argentina respectively have total scores between 0 and 16 points, while 74.1% and 73.8% of students from England and USA respectively have total scores between 16 and 32 points.

**Score Distributions, Argentina and Colombia**

![Graph showing score distributions for Argentina and Colombia](Figure 1. Argentina and Colombia Score Distributions)

**Score Distributions, England and USA**

![Graph showing score distributions for England and USA](Figure 2. England and USA Score Distributions)
Item Level Analyses

This study used two DIF methods to examine item level equivalence, OLR and Poly-SIBTEST. Two methods were chosen because it has been shown that different DIF detection methods perform best under different conditions, so it is advisable to consider items as containing DIF if they are identified by two methods (Ercikan et al., 2004; Kim & Cohen, 1995). Results of these two analyses are presented next, followed by a comparison of these results.

*Ordinal Logistic Regression*

Results of the OLR analysis are shown in Table 7. OLR identified between 54 and 75% of the items in the Reader booklet as being differentially functioning. All items identified as DIF by OLR were classified as having negligible effect sizes using Jodoin and Gierl's (2001) effect size classification scheme outlined in the previous chapter, indicating that although significant the amount of variance accounted for by the grouping variable (individual country versus international data) is small.

Table 7

*Ordinal Logistic Regression DIF Results*

<table>
<thead>
<tr>
<th>DIF Items</th>
<th>Argentina</th>
<th>Colombia</th>
<th>England</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>13</td>
<td>18</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Percentage</td>
<td>54.2</td>
<td>75.0</td>
<td>66.7</td>
<td>66.7</td>
</tr>
</tbody>
</table>

Note: Total number of items = 24

Note: p ≤ .01
Results of the standard, one-item-at-a-time, DIF analysis using Poly-SIBTEST also reveal a large percentage of DIF items, as shown in Table 8. Table 8 shows that the DIF items identified in each country are approximately balanced in terms of whether the focal group (country) or the reference group (international) is favoured. Using the effect size classification of Roussos and Stout (1996), Poly-SIBTEST identified items at all three effect size levels (negligible, moderate and large).

Table 8

<table>
<thead>
<tr>
<th>Country</th>
<th>Argentina</th>
<th>Colombia</th>
<th>England</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIF Favours</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Items</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>% of Items</td>
<td>41.7</td>
<td>54.2</td>
<td>54.2</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Notes: Total number of items = 24

F = Focal Group, R = Reference Group

p ≤ .01

Comparison of OLR and Poly-SIBTEST DIF Results

In this study, the results of DIF analyses using OLR and Poly-SIBTEST are compared for two purposes. The first is as a method of verifying DIF identification, as noted previously, and to seek convergence between the two DIF methods to identify bundles of items for further analysis. The second purpose is to add to the few existing empirical studies in which such a comparison is made.
A consideration of 24 items in each of 4 countries provides 96 comparison points of potential agreement between the two DIF detection methods. Over these 96 comparison points there was 77% agreement between the two methods as to the presence or absence of DIF at the significance level of $p \leq .01$.

Although there was high agreement between the two methods as to the presence or absence of DIF, there were differences in the effect sizes detected by the two methods. As noted above, all items identified as DIF by OLR were classified as negligible, whereas Poly-SIBTEST identified items at all three classification levels of effect size. Effect size classifications of both methods are shown in Table 9.

Table 9

<table>
<thead>
<tr>
<th>Country</th>
<th>Negligible</th>
<th>Moderate</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Poly-OLR</td>
<td>Poly-OLR</td>
<td>Poly-OLR</td>
</tr>
<tr>
<td>ARG</td>
<td>13</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>COL</td>
<td>18</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>ENG</td>
<td>16</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>USA</td>
<td>16</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: $p \leq .01$

Items which were identified as DIF at $p \leq .01$ by both methods are shown in Table 10. As can be seen, all 24 items were identified as DIF by both methods in at least one of the four countries. Since the agreement level between the
methods was so high with all items being identified as DIF, further examination
was carried out only on items that were identified by both methods in at least three
of the four countries. Seven such items were found, being items 4, 6, 8, 9, 15, 18
and 22. These seven items provided the basis on which bundles were chosen for
the differential bundle study.

Table 10

*Items Identified as DIF by Ordinal Logistic Regression and Poly-SIBTEST*

<table>
<thead>
<tr>
<th>Item #</th>
<th>Arg</th>
<th>Col</th>
<th>Eng</th>
<th>Usa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>15</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>17</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>22</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Bundle Level Analysis

The seven items which were identified as containing DIF by both OLR and Poly-SIBTEST in at least three of the four countries were studied further in an attempt to understand whether there may be some inherent similarities among the items on which to form an organizing principle for bundling. The items were analyzed according to their specifications in the PIRLS 2001 data files. Three of the 7 items were multiple choice and 4 were constructed response items. Four of the items were intended to assess the purpose of reading for literary experience and 3 were intended to assess acquiring and using information. Thus item type and purpose of reading were not considered to be potential sources of DIF. Of interest though, six of the 7 items (items 6, 8, 9, 15, 18 and 22) were designed to assess the reading comprehension process of interpreting and integrating ideas and information, and the remaining one item (item 4) was designed to assess the focus and retrieve reading comprehension process.

The finding that 6 of the 7 items were intended to assess the same reading comprehension process provided an organizing principle on which to bundle items for the DBF analysis. Based on Shealy and Stout’s (1993) multidimensional model for DIF, the reading comprehension cognitive levels were hypothesized to contain a secondary dimension which represents a potential source of differential functioning. All Reader booklet items from the reading comprehension cognitive process of interpreting and integrating ideas and information were analyzed as a suspect bundle using Poly-SIBTEST. All remaining items in the Reader booklet formed the matching subtest items for this analysis. Results revealed significant
DBF against both Argentina and Colombia, when the basis for bundle formation is the reading comprehension cognitive level of interpreting and integrating ideas and information. At the same time, for this cognitive level bundle, significant DBF in favour of the USA was identified. Similarities amongst the items in this cognitive level were noted when reviewing the items. Four items in this bundle required explanations, six of the items used the word "how", and four of the items were on the topic of the characters' feelings. Two examples of items in this bundle are: item 9, how did the hare's feelings change during the story?; and item 23, write two different feelings Halla might have after she has set the pufflings free and explain why she might have each feeling.

Items belonging to the lower cognitive level focus and retrieve reading comprehension process were then analyzed as a suspect bundle, with all other Reader items forming the matching subtest. For this grouping of lower cognitive level items, significant DBF was identified in favour of Argentina and Colombia, and against England and the USA. Items in this bundle tend to begin with the words "what", "where" or "why". Two examples are: item 12, why are puffins clumsy at takeoffs and landings?; and item 16, what happens during the nights of the pufflings?

DBF analyses results for both bundles of items are shown in Table 11. A positive $\beta$ estimate together with a significant $p$ value ($\leq 0.05$) indicates differential bundle functioning against the focal country (favouring the reference group), and a negative $\beta$ estimate together with a significant $p$ value indicates differential bundle functioning against the reference group and favouring the focal country.
group. As shown, there is significant DBF against Argentina and Colombia in the “interpret and integrate” bundle, and in favour of the USA. In the “focus and retrieve” bundle there is significant DBF favouring Argentina and Colombia and against England and the USA.

Table 11

_Differential Bundle Functioning Analyses_

<table>
<thead>
<tr>
<th>Suspect Bundle</th>
<th># Items</th>
<th>Focal Group</th>
<th>Estimate</th>
<th>p-value</th>
<th>DBF against (group):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpret &amp; Integrate</td>
<td>12</td>
<td>ARG</td>
<td>0.71</td>
<td>0.00</td>
<td>Focal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COL</td>
<td>0.71</td>
<td>0.00</td>
<td>Focal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ENG</td>
<td>0.14</td>
<td>0.19</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>USA</td>
<td>-0.55</td>
<td>0.00</td>
<td>Reference</td>
</tr>
<tr>
<td>Focus &amp; Retrieve</td>
<td>8</td>
<td>ARG</td>
<td>-0.164</td>
<td>0.02</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>COL</td>
<td>-0.333</td>
<td>0.00</td>
<td>Reference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ENG</td>
<td>0.121</td>
<td>0.04</td>
<td>Focal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>USA</td>
<td>0.317</td>
<td>0.00</td>
<td>Focal</td>
</tr>
</tbody>
</table>

Given that each suspect bundle contained more than 5 items, no classification guidelines exist for interpreting effect sizes for these results (Boughton et al, 2000; Gierl et al, 2003; Gierl & Bolt, 2001). However, in all cases where significant DBF was identified, the absolute values of the β-estimates were greater than 0.1 which would have been classified as moderate to large had the bundles contained fewer items.
Scale Level Analyses

In order to investigate the degree of scale level equivalence, and whether
differential item and bundle functioning are manifested at the test level, an
analysis of DTF was conducted using PARDUX (Burket, 1991). Item response
theory parameters were estimated for the international reference sample and for
each of the four focal country samples separately using the 2PL and PPC models
(Fitzpatrick et al., 1996) discussed in chapter 3. In order to link item parameters
from separate country calibrations to their respective international sample a robust
mean and sigma transformation as described by Stocking and Lord (1983) was
carried out. All items were included in the transformation. Table 12 shows the
average discrimination and difficulty item parameters in the original metric prior
to transformation (columns “a” and “b”), the discrimination and difficulty of the
items in the transformed metric (columns “a(t)” and “b(t)”) and the transformation
parameters used (columns “M1” and “M2”).
An analysis of the correlations between country specific and international item parameters was conducted. This provides information about the similarity of the ordering of discrimination and difficulty parameters in the two sets of calibrations. These correlations are summarized in Table 13. The discrimination \((a)\) parameter correlations between country specific calibrations and the international calibration ranged from 0.66 (for Argentina) to 0.85 (for Colombia). The correlations of the difficulty \((b)\) parameters were higher. They ranged from 0.89 (for Argentina) to 0.94 (for USA). These correlations tend to be moderate to high compared to other item parameter correlations conducted with international assessments (Ercikan & Koh, 2005).

### Table 12

*Average original and transformed discrimination and difficulty parameters*

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>a</th>
<th>b</th>
<th>a(t)</th>
<th>b(t)</th>
<th>M1(^a)</th>
<th>M2(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>24</td>
<td>1.16</td>
<td>0.51</td>
<td>1.14</td>
<td>-0.34</td>
<td>1.01</td>
<td>-0.85</td>
</tr>
<tr>
<td>COL</td>
<td>24</td>
<td>1.04</td>
<td>0.28</td>
<td>1.36</td>
<td>-0.49</td>
<td>0.77</td>
<td>-0.71</td>
</tr>
<tr>
<td>ENG</td>
<td>24</td>
<td>1.32</td>
<td>-1.26</td>
<td>1.30</td>
<td>-0.75</td>
<td>1.01</td>
<td>0.52</td>
</tr>
<tr>
<td>USA</td>
<td>24</td>
<td>1.26</td>
<td>-1.17</td>
<td>1.35</td>
<td>-0.66</td>
<td>0.93</td>
<td>0.43</td>
</tr>
</tbody>
</table>

\(^a\)M1: Stocking and Lord multiplicative factor

\(^b\)M2: Stocking and Lord additive factor
Table 13.

Correlation of $a$ and $b$ parameters for each
country with the international parameters

<table>
<thead>
<tr>
<th>Country</th>
<th>$a$ parameters</th>
<th>$b$ parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>0.66</td>
<td>0.89</td>
</tr>
<tr>
<td>Colombia</td>
<td>0.85</td>
<td>0.92</td>
</tr>
<tr>
<td>England</td>
<td>0.71</td>
<td>0.93</td>
</tr>
<tr>
<td>USA</td>
<td>0.84</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Test characteristic curves were computed and graphed using MODFIT (Stark, 2001) based on the international and linked country parameters. TCC comparisons for each country are shown in Figures 3 to 6. Each country's TCC looks very similar to its international TCC. These results tend to support the notion of score comparability at the test level (i.e., DTF is not evident in the TCC comparisons).
Figure 3. TCC Comparison for Argentina

Figure 4. TCC Comparison for Colombia
Figure 5. TCC comparison for England

Figure 6. TCC comparison for USA
Differences in the percent of maximum score estimated between the international sample and individual countries were determined together with descriptive statistics of the differences. These results are shown in Table 14. The negative values for differences for England and the USA indicate that these two countries would receive higher scores if scores were created based on their own country parameters instead of on international parameters. The positive differences for Argentina and Colombia indicate they receive higher scores based on the international parameters than they would if scores were based on their own country parameters.

Table 14.

Descriptive statistics of differences between country TCCs and international TCCs

<table>
<thead>
<tr>
<th>Country</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>-0.0084</td>
<td>0.0349</td>
<td>0.0175</td>
<td>0.0129</td>
</tr>
<tr>
<td>Colombia</td>
<td>-0.0118</td>
<td>0.0276</td>
<td>0.0066</td>
<td>0.0124</td>
</tr>
<tr>
<td>England</td>
<td>-0.0585</td>
<td>0.0224</td>
<td>-0.0202</td>
<td>0.0300</td>
</tr>
<tr>
<td>USA</td>
<td>-0.0429</td>
<td>0.0146</td>
<td>-0.0118</td>
<td>0.0193</td>
</tr>
</tbody>
</table>

Differences between focal and reference TCCs were examined along the ability scale, and results are plotted in Figures 7 and 8. A comparison of the score distributions shown in Figures 1 and 2 with the TCC differences shown in Figures 7 and 8 reveal that the region along the ability scale where Argentina and Colombia would be most advantaged by the use of international parameters coincides with the region in the score distribution where most students are located.
Conversely, the region along the ability scale where England and USA would be disadvantaged coincides with the region in the score distribution where the fewest students are found in those two countries.

**Figure 7.** Argentina and Colombia, Differences in TCCs based on international versus country parameters.

**Figure 8.** England and USA, Differences in TCCs based on international versus country parameters.
Model-Data Fit

An evaluation of model-data fit using Yen's (1984) Q1 fit statistic was also conducted using PARDUX (Burket, 1991) to test the goodness of fit between the IRT models and test data. Items were considered to fit poorly if \(|z| \geq 4.0\) for individual country data and if \(|z| \geq 18.0\) for the international samples. Four of the 24 items were found to display poor fit. Item #19 exhibited poor fit across all four countries and all four international data sets. Item #17 exhibited poor fit in all four international data sets, as well as in Argentina and Colombia. Items #7 and #18 displayed poor fit in Colombia and Argentina respectively. Items displaying poor fit are summarized in Table 15. Items 17, 18 and 19 are items that were identified as DIF using both DIF methods in either 2 (items 17 and 19) or 3 (item 18) countries, and all three items are intended to assess the reading comprehension process of interpreting and integrating ideas and information. They also contributed to the DBF which was observed when the interpret and integrate reading comprehension process items were analyzed as the suspect bundle. Item 7 was identified as DIF by both DIF methods in one country (Colombia), and is intended to assess the reading comprehension process of making straightforward inferences.
Evidence of construct comparability was analyzed through an examination of factor structures for each of the four countries separately using exploratory factor analysis (EFA). The method of extraction was Principal Components Analysis, using LISREL/PRELIS, version 2.50 (SSI Scientific Software International (2000). The criteria for establishing factor structure used in this study were: the Kaiser-Guttman rule, eigenvalue ratios of factor 1 to factor 2, and examination of scree plots. Using the Kaiser-Guttman rule for the lower bound, the number of factors is equal to the number of components with eigenvalues that are greater than or equal to 1.0 (Gorsuch, 1983). The factor analysis results for factors with eigenvalues greater than 1.0 are shown in Table 16.

**Table 15.**

*Q1 Goodness of fit results: items identified as misfit*

<table>
<thead>
<tr>
<th>Item No.</th>
<th>ARG</th>
<th>COL</th>
<th>ENG</th>
<th>USA</th>
<th>ARG</th>
<th>COL</th>
<th>ENG</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>6.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>4.04</td>
<td>4.33</td>
<td></td>
<td></td>
<td>38.25</td>
<td>37.58</td>
<td>39.77</td>
<td>37.21</td>
</tr>
<tr>
<td>18</td>
<td>4.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>6.56</td>
<td>5.38</td>
<td>10.06</td>
<td>12.58</td>
<td>63.54</td>
<td>70.54</td>
<td>63.17</td>
<td>61.84</td>
</tr>
</tbody>
</table>

\[^a|z| \geq 4.0\] \[^b|z| \geq 18.0\]
Table 16

**PCA Eigenvalues and variance explained for each factor**

<table>
<thead>
<tr>
<th>Country</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Ratio of Factor 1 to Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>8.68</td>
<td>1.72</td>
<td>1.38</td>
<td>1.19</td>
<td>1.11</td>
<td>1.04</td>
<td>5.04</td>
</tr>
<tr>
<td>% Variance</td>
<td>36.15</td>
<td>7.16</td>
<td>5.76</td>
<td>4.96</td>
<td>4.63</td>
<td>4.32</td>
<td></td>
</tr>
<tr>
<td>Cumulative %</td>
<td>36.15</td>
<td>43.31</td>
<td>49.07</td>
<td>54.03</td>
<td>58.66</td>
<td>62.99</td>
<td></td>
</tr>
<tr>
<td>COL</td>
<td>7.96</td>
<td>1.48</td>
<td>1.41</td>
<td>1.11</td>
<td>1.02</td>
<td>-</td>
<td>5.38</td>
</tr>
<tr>
<td>% Variance</td>
<td>33.16</td>
<td>6.19</td>
<td>5.89</td>
<td>4.64</td>
<td>4.24</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Cumulative %</td>
<td>33.16</td>
<td>39.35</td>
<td>45.24</td>
<td>49.88</td>
<td>54.12</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ENG</td>
<td>10.05</td>
<td>1.59</td>
<td>1.12</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
<td>6.32</td>
</tr>
<tr>
<td>% Variance</td>
<td>41.88</td>
<td>6.62</td>
<td>4.66</td>
<td>4.2</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Cumulative %</td>
<td>41.88</td>
<td>48.5</td>
<td>53.16</td>
<td>57.36</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>9.82</td>
<td>1.57</td>
<td>1.23</td>
<td>1.03</td>
<td>-</td>
<td>-</td>
<td>6.25</td>
</tr>
<tr>
<td>% Variance</td>
<td>40.92</td>
<td>6.53</td>
<td>5.14</td>
<td>4.3</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Cumulative %</td>
<td>40.92</td>
<td>47.45</td>
<td>52.6</td>
<td>56.9</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 16 presents eigenvalues, percentage of variance accounted for by each factor solution, the cumulative percent of variance accounted for by each additional factor, and the eigenvalue ratios of factor 1 to factor 2, for all four countries. Six factors represent the data for Argentina, 5 for Colombia, and 4 for England and the US. The ratio of eigenvalues for factor 1 to factor 2 range from...
5.04 to 6.32, providing evidence of one main factor underlying the data, as well as support for the IRT assumption of essential unidimensionality of the Reader items in all four countries. The one main factor accounts for approximately 33% to 42% of the variance in the data while the second factor accounts for between 6 and 7% of the variance.

Scree plots were also examined in this study. Scree plots are graphs of eigenvalues plotted against each factor representing the data. The number of factors that precede the bend, or elbow, in the plot indicates the maximum number of factors that can be extracted from the data. The scree plots, shown in Figure 9, show that the data are represented by one major factor for all four countries, evidenced by the sharp bend in each graph at the second factor.

**Scree Plots**

![Scree Plots](image)

*Figure 9. Scree plots for Reader booklet data.*
Summary

This study presents a score scale comparability analysis of the PIRLS 2001 Reader booklet data for the countries of Argentina, Colombia, England and USA at the item, bundle, and test levels of analysis. The procedures used and the study results were presented in this chapter.

Large differences were found in the raw score means between individual countries and their respective reference group samples, with Argentina and Colombia scoring 0.78 and 0.74 standardized score difference units below their reference means, while England and the USA score 0.47 and 0.41 standardized score units above their reference means. An examination of score distributions for the four countries revealed that nearly three-quarters of students from Argentina and Colombia score in the lower half of the score range, while nearly three-quarters of students from England and USA score in the upper half of the score distribution.

Exploratory factor analysis results indicate that the overall test structure is very similar for all four countries, with one main factor underlying the data. However, secondary factors with eigenvalues greater than 1.0 are present in all countries, and these factors may give rise to differential functioning between the groups.

At the item level of analysis, both OLR and Poly-SIBTEST analyses identified a large percentage of differentially functioning items in all four countries. These results are consistent with previous studies of large scale educational assessments. There was a high degree of agreement between the two
DIF detection methods as to the presence or absence of DIF. However, the effect size classifications resulting from the two methods were different, with OLR identifying all items as negligible, whereas Poly-SIBTEST identified approximately 20% of the DIF items as negligible, 30% as moderate, and 50% as large effect size. Due to the large percentage of DIF items identified in the data, inclusion criteria were set for the DBF study. Only those items that were identified as DIF by both methods in three or four countries were included for DBF analysis. Seven items met the criteria for further inclusion.

The analysis of item content specifications to seek out organizing principles on which to form hypotheses for the DBF analysis revealed that six of the seven DIF items were intended to assess the higher cognitive level reading comprehension process of integrating and interpreting ideas and information and the remaining item was intended to assess the lower cognitive level reading comprehension process of focusing and retrieving specific information.

Two DBF analyses were conducted, one in which the suspect items were comprised of all items from the interpret and integrate process, and one in which the suspect items were comprised of all items from the focus and retrieve process. Significant DBF was revealed against Argentina and Colombia, and in favour of USA, for the higher cognitive level reading comprehension process items. Significant DBF in favour of Argentina and Colombia, and against England and USA, was revealed for the lower cognitive level process items.

At the test level, item response theory was used to create test characteristic curves for comparison between country and international (reference) groups.
There was a moderate to high correlation between country and international parameters, indicating similar ordering of the parameters in the two sets of calibrations. Analysis of model-data fit revealed four items with poor fit to the model, three of which belong to the interpret and integrate reading comprehension process. Test characteristic curves for each country and its international reference sample are very similar, indicating possible score comparability, or lack of differential functioning, at the test level. An analysis of differences between international and country TCCs along the latent ability continuum reveals that Argentina and Colombia may be favoured by the use of international parameters to create country scores in the region where these countries have the greatest score density, that is, in the lower half of the score distribution. Conversely, England and the USA are disadvantaged by the use of international parameters in the region where they have fewer numbers of students.

The following chapter will provide a discussion of these results, together with implications for practice, contributions of these findings to the research literature, future directions for study, and limitations of this study.
CHAPTER V: DISCUSSION

Large-scale international assessments of achievement necessitate the translation and adaptation of one assessment into many different languages for use in a wide variety of countries and cultures worldwide. Results from these assessments are used to make comparisons across countries, which require that scores be comparable across countries. Evidence of score comparability is required to make meaningful comparisons across countries.

This study investigated the degree of score comparability in the Reader booklet of the PIRLS 2001 large-scale educational assessment, using results from four countries: Argentina, Colombia, England and USA. Four research questions were addressed:

1. What is the degree of DIF in the PIRLS data? What are similarities in DIF detection by OLR and Poly-SIBTEST?

2. At the bundle level, are there any bundles of items in the PIRLS Reader booklet that may result in DIF amplification or cancellation?

3. At the test level, what are similarities or differences in test characteristic curves based on the international parameters versus those based on individual country parameters? Are the international parameters appropriate for creating country scores?

4. Why do large degrees of differential functioning in large-scale assessments at the item level result in only small degrees of differences at the test level?
The findings from this study confirm current DIF research in the field of score comparability in large-scale assessment, and expand on the research base regarding DBF and comparability of test characteristic curves at the test level.

This chapter provides a discussion of the research results, together with implications for practice, contributions of these findings to the research literature, future directions for study, and limitations of this study.

Discussion of Findings

Research Question #1: Differential Item Functioning Analyses

Differential item functioning analyses were conducted using two methods, OLR and Poly-SIBTEST, to investigate score comparability in the PIRLS 2001 Reader booklet at the level of individual items. Both methods identified large percentages of DIF against all four countries when compared to their international reference sample. Converging the results of the two methods to verify the presence of DIF showed that all 24 items were identified as DIF by both methods in at least one country. In addition, 11 items (45.8%) were identified by both methods in 2 countries, 4 items (16.67%) were identified in 3 countries, and 3 items (12.5%) were identified as DIF in all 4 countries by both methods.

The results of this study confirm many previous studies that have demonstrated a large degree of DIF in international large-scale assessments (Ercikan et al., 2004; Sireci & Allalouf, 2003). When comparisons at the level of individual items are made between an individual country and a sample intended to represent all countries, there are a significant number of items that function differentially between these groups. Despite being matched on ability, the groups
have different probabilities of correctly responding to individual items. The DIF that is present is fairly evenly distributed across the four countries involved, and does not appear to systematically favour focal or reference groups. This indicates that language of testing alone does not account for the DIF that is present.

Recall from Chapter 1 that the finding of empirical evidence of DIF is necessary but not sufficient to conclude bias is present. Once an item has been determined to exhibit DIF, it is necessary to attempt to identify sources of the DIF and use judgment to decide whether the item is biased.

One strategy for analyzing possible sources of DIF is to examine the content and topic of DIF items (Ercikan, 2002). The 7 items that were identified as DIF by both methods in three or four countries were examined for their content. Six of these 7 items were intended to assess one cognitive level of reading comprehension, the level of interpreting and integrating ideas and information from the text. As a result of these analyses, a hypothesis was formed that the bundle of items intended to assess the cognitive level of interpreting and integrating ideas and information would contain DBF. This hypothesis was tested using Poly-SIBTEST in a DBF analysis, the results of which are discussed in a following section which addresses the findings of the DBF study.

Comparison of Results from Poly-SIBTEST and Ordinal Logistic Regression

This study has demonstrated a high correspondence in results between the two DIF methods with 77% agreement between them as to the presence or absence of DIF. This was expected due to previous research that has shown that each of
the two methods is capable of identifying DIF in large-scale assessments in circumstances such as those in this research study.

Once DIF has been identified it is important to determine an effect size to aid in understanding whether statistically significant DIF has practical significance or not. In this study, all items identified by OLR were identified as negligible according to the guidelines established by Jodoin and Gierl (2001). Recall that Jodoin and Gierl used cubic curve regression to predict the OLR effect size from the SIBTEST effect size, and determined a concordance for negligible, moderate and large effect sizes between the two methods. However, while all OLR DIF items were classified as negligible, only 19% of the items identified as DIF by Poly-SIBTEST were classified as negligible. In addition, Poly-SIBTEST classified 31% and 50% of the items as moderate and large DIF effect sizes respectively. Therefore, this research does not support the findings of Jodoin and Gierl on the correspondence between the effect sizes for the two methods.

Research Question #2: Differential Bundle Functioning Analyses

DBF analyses have greater power than analyses of individual items. When bundles of items are selected using an organizing principle, and the substantive characteristics of the items are revealed in the structure of the data, differential functioning may be attributed to the basis on which the items are organized into bundles. Based on the joint results of the two DIF analyses, item type, purpose for reading, and language of testing were hypothesized not to be sources of DIF. However, it was hypothesized that reading comprehension cognitive level may constitute a source of differential functioning. Sources of bias that may contribute
to reading comprehension differential functioning in this study were also considered. One source of bias that may be particularly relevant to the results of this study is bias that is introduced when test content is not equally aligned with the curricula for all groups being tested (The Standards, AERA, APA & NCME, 1999). Differential functioning in reading comprehension items across countries in PIRLS 2001 Reader booklet may reflect bias resulting from differences between curricula, what is intended to be taught or what is actually taught in the participating countries. Other potential sources of bias that may be relevant to the results of this study are opportunity to learn and motivation related to differential access to learning materials, and culture (in that students from different cultures may approach reading comprehension differently).

In order to test the hypothesis of differential functioning related to reading comprehension cognitive levels, all items in the PIRLS 2001 Reader booklet from the reading comprehension process of interpreting and integrating ideas and information were analyzed for DBF. Indeed, significant DBF was found against Argentina and Colombia, and in favour of the USA for “the interpret and integrate” bundle. Next, all items from the lower cognitive level “focus and retrieve” reading comprehension items were analyzed as a bundle, and significant DBF was found in favour of Argentina and Colombia, and against England and USA.

These results support the hypothesis that reading comprehension cognitive level is a source of differential functioning in the PIRLS 2001 Reader booklet, and provide evidence of lack of score comparability at the bundle level of analysis.
There is a considerable amount of evidence in the PIRLS 2001 documentation that supports the hypothesis that reading comprehension DIF represents a source of bias caused by differences in curricula, and opportunity to learn across the four countries as well as differences in motivation of the students. The PIRLS 2001 Encyclopedia (Mullis, Martin, Kennedy, & Flaherty, 2002) provides a description of the participating countries’ policies and practices that guide school organization and classroom reading instruction in the lower grades and provides contextual information necessary for appropriate interpretation of PIRLS results. A review of the Encyclopedia reveals potentially significant differences between the four countries in this study related to curriculum and differential opportunity to learn the skills associated with reading comprehension. Examples are differences in: the age at which reading instruction begins and the percentages of students receiving early reading instruction; the number of years of formal education students have had when they are assessed; the curriculum taught as well as differences between the intended curriculum and what is actually taught; and student access to interesting reading materials both at home and at school (Arrieta & Parreño, 2002; Rush & Kapinus, 2002; Sáenz, Rocha, Garcia, & Riveros, 2002; Twist, 2002).

The availability of books is related to student motivation to learn to read and enjoy reading. Students are motivated to read by the availability of a variety of informational and literary resources across a wide range of difficulties (Early Reading Expert Panel, 2003; Guthrie & Alao, 1997). Arrieta and Parreño (2002) state that the only book available to each Argentinean student is a graded reader or
his/her own textbook. In fact, “most classes (and even schools) do not have a bookshelf or reading corner for students to use” (pp. 4-5). Argentinean students are disadvantaged by a lack of reading materials which represents a source of bias related to motivation.

Several studies have been noted in which culture is shown to affect interpretation of what is read (see, for example, Bock, 2006; Murata, 2007; Reynolds, Taylor, Steffensen, Shirey, & Anderson, 1982). Bock demonstrated that the interpretation of what is read varies according to ethnicity or culture independent of reading comprehension ability. Li, Cohen and Ibarra (2004) contend that cultural context can confound measurement of knowledge on standardized tests, and that culturally distinct groups have specific patterns of thinking and learning that may constitute sources of DIF.

Research Question #3: Differential Test Functioning and TCC Comparisons

Results of DTF analyses using TCC comparisons are similar to those seen in previous research (Ercikan & Sandilands, in press; Ercikan & Gonzalez, 2008). These findings appear to support score comparability at the test level between country and international TCCs with a maximum absolute value difference of approximately 6%.

One additional contribution of this study is that it extends previous research by examining TCC differences along the ability scale and comparing each country’s TCC difference to its score distribution. Results reveal that the region along the ability scale where Argentina and Colombia would be most advantaged by the use of international parameters to create their country scores
coincides with the region in the score distribution where most of their students are located. Therefore, the scores of a large proportion of students in Argentina and Colombia may be artificially inflated (up to 3.5% of maximum score) through the use of international item parameters to create individual country scores. Conversely, the region along the ability scale where England and USA would be disadvantaged coincides with the region in the score distribution where the fewest students are found in those two countries. In other words, a relatively small proportion of student scores in England and USA may be artificially deflated (by up to 6% of maximum score for England and 4% for USA) through the use of international item parameters. It therefore is possible that the use of international item parameters to create individual country scores may have provided a relative advantage to Argentina and Colombia, which creates the appearance of comparability at the test level (a very small degree of DTF is evident in the TCCs) despite incomparability at the item level (high levels of DIF) and at the bundle level (DBF based on cognitive levels in reading comprehension processes).

In addition, analysis of model-data fit revealed four items (17%) with poor fit to the IRT model, indicating that for these items, the item response theory assumption of model fit to data may not be met. A poorly fitting IRT model will not yield invariant item or ability parameters. In other words, one of the fundamental assumptions of IRT will not be met: that the item parameters not depend on the ability distribution of the examinees, and the examinee parameter not depend on the set of test items (Hambleton et al., 1991). Three of the poorly-fitting items are from the interpret and integrate reading comprehension cognitive
level, which lends supporting evidence to the findings that these cognitive level items may be functioning differentially in the PIRLS 2001 Reader booklet.

Research Question #4: Why do large degrees of differential functioning in large-scale assessments at the item level result in only small degrees of differences at the test level?

This fourth research question was addressed by examining the results of the analyses for the first three research questions. Similar to other research, the TCC comparisons appear to support score comparability at the test level; yet both DIF and DBF have been found in the Reader booklet data which fails to provide support for score comparability at the item and bundle levels. Three explanations seem feasible for this apparent discrepancy. First is the possibility that bundles of items are cancelling each other at the test level in a manner similar to which Nandakumar (1993) had observed with item cancellation. The DBF of higher cognitive level items may cancel that of the lower cognitive level items in such a way that at the test level neither is detected. The second possibility is that the current method of detecting DTF through examination of IRT TCCs is not sensitive enough to capture potentially subtle amounts of differential functioning which are spread over the ability scale. A third possibility, and one that seems to be supported by the findings of this study, is that the current scaling methods that use IRT to estimate item parameters based on an international sample are not appropriate for creating individual country scores.
Implications, Contribution to Research and Further Research Suggestions

Each of the research findings noted in the previous sections has implications for practice, and leads to suggestions for further research. These implications will have an impact on decisions by many stakeholders in the realms of test development and use.

The first implication is related to validity and the use of scores from the PIRLS 2001 Reader booklet to compare country results. This study has added to the research by providing statistical evidence of differential functioning at the item and bundle levels for items that tap higher cognitive level reading comprehension processes, specifically interpretation and integration of information and ideas from what is read. Substantive evidence to support this study’s statistical finding of differential functioning has been found in the literature on reading comprehension, providing support that the differential functioning is indicative of bias. Recall that equivalence can be defined as a lack of bias which is required in order to make valid comparisons across cultures or other groups (van de Vijver & Tanzer, 2004). Interpretations of score comparisons across countries that took part in PIRLS 2001 should therefore be made with caution as, without appropriate validity evidence, such comparisons may lack meaning and may result in erroneous decisions. Test developers and publishers have a duty according to the Standards (AERA, APA & NCME, 1999) and the ITC Guidelines (International Test Commission, 2000) to establish equivalence through the application of appropriate statistical techniques and to provide validity evidence for all intended populations. The results of this
study point to the need for test developers and publishers to fulfill these obligations prior to assessment results being used for important policy or other high-stakes decisions. It may be necessary for test developers to provide evidence at all three levels of analysis in order to allow test users to make appropriately informed decisions.

Further implications arise from the discrepancy between score comparability evidence at the test level versus the item and bundle levels. This may not represent the absence of differential functioning at the test level, but rather may reflect bundle cancellation, insensitivity of current methods to detect DTF, or the use of inappropriate methods of determining country scores based on international item parameters. If DTF is present but not being detected, true group differences will be masked and test results may be used to make unfounded decisions such as policy, curriculum or funding decisions. Further research is needed to gain a better understanding as to which, if any, of these three postulated factors is contributing to the apparent discrepancy between item and test level analyses. In addition, new methods of detecting DTF may be necessary to increase the likelihood of detecting DTF when it is present.

This study has contributed to research by introducing preliminary evidence for the possibility that the use of international item parameters to create individual country scores may provide a relative advantage to some countries due to the locations of their score distributions. This relative advantage may then create the appearance of comparability at the test level despite evidence for incomparability at both the item and bundle levels. Significant research indicates that when there is
insufficient evidence to support measurement invariance across groups, score scales should be based on separate country calibrations which are then linked to establish comparability (Cook, 2006; Ercikan & Gonzalez, 2008; Sireci, 1997). The implication for test developers may be that score scales should be estimated separately by country and then linked. Further research should be conducted in this regard. Although this study has demonstrated the possibility of a relative advantage to some countries through the use of international parameters, it has not established whether there is or may be a significant effect on country comparisons or ordering of countries due to such advantage. Future studies should attempt to quantify the relative advantages and disadvantages to determine their overall effect.

This study also has implications regarding comparison of effect sizes from OLR and Poly-SIBTEST. Based on statistical significance and effect sizes, test developers and administrators make decisions to exclude items from assessments at the test development stage, or from scoring after a test has been administered. Clearly, in this study, different interpretations regarding practical significance and decisions would flow from the same data, depending on which method was used to analyze the data and determine the effect size. Thus the comparability of the existing guidelines for determining effect sizes with these two DIF methods comes into question. Jodoin and Gierl’s (2001) effect size prediction was based on equal sample sizes for the focal and reference groups, whereas this study may be limited by the use of unequal sample sizes. It is important to note, though, that the presence of unequal sample sizes is a common occurrence in DIF analyses of
large-scale assessments and therefore should be taken into account in setting effect size criteria. Jodoin and Gierl also found in their power simulation study that OLR power tended to be lower in the condition of unequal group size combined with unequal ability; again this condition exists in the current study. Future research may be required to investigate the effect of unequal sample size and ability on effect size criterion and comparability across the two DIF detection methods.

This study has contributed to research in the field of large-scale educational assessments such as PIRLS by demonstrating the use of PolySIBTEST DBF analyses to establish potential causes for differential functioning seen at the item level. Using statistical results from the DIF study, hypotheses were generated as to the cause of DIF and tested in the DBF study. Results of the DBF study confirmed DIF related to cognitive levels of reading comprehension.

Lastly, although this study revealed potential sources of bias in the PIRLS assessment, there are additional sources of bias that could be contributing to DIF. An area requiring further study is the effect of age differences and developmental abilities on reading comprehension results. Students from England and USA represent a fairly homogeneous age group, whereas the ages of students in Argentina and Colombia exhibit much greater variance. This study has not addressed the impact that age variation within country may have on PIRLS 2001 comparability results.
Limitations of this Study

This study has examined score comparability in international educational assessments using data from the PIRLS 2001 Reader booklet. It is unknown whether these results would generalize to other international educational assessments, or to the remaining booklets in the PIRLS 2001 study. Further studies should be conducted to determine whether similar results would be found in other international educational assessments.

In this study an international sample reference group was created with which to compare the country focal groups, which resulted in unequal reference and focal group sample sizes. The sample sizes for the reference groups were 6800 students whereas the country sample sizes ranged from 777 to 1285. The results of this study should be replicated using more equally balanced sample sizes.

Summary

This study analyzed score scale comparability in the PIRLS 2001 Reader booklet data in four countries. Score comparability was found to be jeopardized by the presence of differential functioning related to the cognitive levels intended to be assessed by the reading comprehension questions. Potential sources of bias may be related to cultural interpretations of reading material, differential curricula and opportunities to learn, and differential motivation to learn reading comprehension across the countries. These findings have significant implications related to the valid use of results from the PIRLS 2001 assessment and the types of decisions which can appropriately be made from these results.
REFERENCES


Stark, S. (2002). MODFIT: Plot theoretical item response functions and examine the fit of dichotomous or polytomous IRT models to response data [computer program]. Department of Psychology, University of Illinois at Urbana-Champaign.


PIRLS Reader

The Natural World

PIRLS
Progress in International Reading Literacy Study
Main Survey 2001
Contents

Part 1

Hare Heralds the Earthquake 1
A short story by Rosalind Kerven

Part 2

Nights of the Pufflings 7
by Bruce McMillan
Hare Heralds the Earthquake

by Rosalind Kerven

There was once a hare who was always worrying. “Oh dear,” he muttered all day long, “oh deary, deary me.”

His greatest worry was that there might be an earthquake. “For if there was,” he said to himself, “whatever would become of me?”

He was feeling particularly anxious about this one morning, when suddenly an enormous fruit fell down from a nearby tree—CRASH!—making the whole earth shake.

The hare leaped up.

“Earthquake!” he cried.

And with that he raced across the fields to warn his cousins.

“Earthquake! Run for your lives!”

All the hares left the fields and madly followed him.

They raced across the plains, through forests and rivers and into the hills warning more cousins as they went.
“Earthquake! Run for your lives!”
All the hares left the rivers and plains, the hills and forests and madly followed.
By the time they reached the mountains, ten thousand hares pounded like thunder up the slopes.
Soon they reached the highest peak. The first hare gazed back to see if the earthquake was coming any closer, but all he could see was a great swarm of speeding hares.
Then he looked in front but all he could see was more mountains and valleys and, far in the distance, the shining blue sea.
As he stood there panting, a lion appeared.
“What’s happening?” he asked.
“Earthquake, earthquake!” babbled all the hares.
“An earthquake?” asked the lion. “Who has seen it? Who has heard it?”
“Ask him, ask him!” cried all the hares, pointing to the first one.
The lion turned to the hare.

“Please Sir,” said the hare shyly, “I was sitting quietly at home when there was a terrible crash and the ground shook and I knew it must be a quake, Sir, so I ran as fast as I could to warn all the others to save their lives.”
The lion looked at the hare from his deep, wise eyes.

“My brother, would you be brave enough to show me where this dreadful disaster happened?”
The hare didn’t really feel brave enough at all, but he felt he could trust the lion.

So, rather timidly, he led the lion back down the mountains and the hills, across the rivers, plains, forests and fields, until at last they were back at his home.

“This is where I heard it, Sir.”
The lion gazed around—and very soon he spotted the enormous fruit which had fallen so noisily from its tree.
He picked it up in his mouth, climbed onto a rock and dropped it back to the ground.

CRASH!
The hare jumped. “Earthquake! Quickly—run away—it’s just happened again!”

But suddenly he realised that the lion was laughing. And then he saw the fruit rocking gently by his feet.

“Oh,” he whispered, “it wasn’t really an earthquake after all, was it?”

“No,” said the lion, “it was not and you had no need to be afraid.”

“What a silly hare I’ve been!”
The lion smiled kindly. “Never mind, little brother. All of us—even I—sometimes fear things we cannot understand.”

And with that he padded back to the ten thousand hares that were still waiting on top of the mountain, to tell them that it was now quite safe to go home.

_The Natural World_
End of Part 1.

Now go to your question booklet.
Every year, black and white birds with orange bills visit the Icelandic island of Heimaey. These birds are called puffins. They are known as "clowns of the sea" because of their bright bills and clumsy movements. Puffins are awkward fliers during takeoffs and landings because they have chunky bodies and short wings.

Halla lives on the island of Heimaey. She searches the sky every day. As she watches from high on a cliff overlooking the sea, she spots her first puffin of the season.

She whispers to herself "Lundi," which means "puffin" in Icelandic.

Soon the sky is speckled with them—puffins, puffins everywhere. They are returning from their winter at sea, returning to Halla's island and the nearby uninhabited islands to lay eggs and raise puffin chicks. These "clowns of the sea" return to the same burrows year after year. It's the only time they come ashore.
Halla and her friends climb over the cliffs to watch the birds. They see pairs tap-tap-tap their beaks together. Each pair they see will soon tend an egg deep inside the cliffs. When the puffin eggs have hatched, the parents will bring fish home to feed their chicks.

Each chick will grow into a young puffling. The nights of the pufflings will come when each puffling takes its first flight. Although the nights of the pufflings are still long weeks away, Halla thinks about getting some cardboard boxes ready.

All summer long the adult puffins fish and tend to their chicks. By August, flowers blanket the burrows. With the flowers in full bloom, Halla knows that the wait for the nights of the pufflings is over.

The hidden chicks have grown into young pufflings. Now it's time for Halla and her friends to get out their boxes and torches for the nights of the pufflings. Starting tonight, and for the next two weeks, the pufflings will be leaving for their winter at sea.

In the darkness of the night, the pufflings leave their burrows for their first flight. It's a short, wing-flapping trip from the high cliffs. Most of the birds splash-land safely in the sea below. But some get confused by the village lights – perhaps they think the lights are moonbeams reflecting on the water. Hundreds of the pufflings crash-land in the village every night. Unable to take off from the flat ground, they run around and try to hide.

Halla and her friends will spend each night searching for stranded pufflings that haven't made it to the water. But the village cats and dogs will be searching, too. Even if the cats and dogs don't get them, the pufflings might get run over by cars or trucks. The children must find the stray pufflings first. By ten o'clock the streets of Heimaey are alive with roaming children.

Halla and her friends race to rescue the pufflings. Armed with torches, they wander through the village, searching dark places. Halla spots a puffling. She races after it, grabs it, and puts it safely in a cardboard box.
For two weeks all the children of Heimaey sleep late in the day so they can stay out at night. They rescue thousands of pufflings.

Every night Halla and her friends take the rescued pufflings home. The next day, with the boxes full of pufflings, Halla and her friends go down to the beach.

It's time to set the pufflings free. Halla releases one first. She holds it up so that it will get used to flapping its wings. Then, holding the puffling snugly in her hands, she swings it up in the air and launches it out over the water beyond the surf. The puffling flutters just a short distance before splash-landing safely.

Day after day Halla's pufflings paddle away, until the nights of the pufflings are over for the year. As she watches the last of the pufflings and adult puffins leave for their winter at sea, Halla bids them farewell until next spring. She wishes them a safe journey as she calls out, "Goodbye, goodbye."

Excerpted from Nights of the Pufflings by Bruce McMillan. © 1995 by Bruce McMillan. Reprinted by permission of Houghton Mifflin Company. All rights reserved.
Directions

In this test, you will read stories or articles and answer questions about what you have read. You may find some parts easy and you may find some parts difficult.

You will be asked to answer different types of questions. Some of the questions will be followed by four choices. You will choose the best answer and fill in the circle next to that answer. Example 1 shows this type of question.

Example 1

1. How many days are there in a week?

   A) 2 days
   B) 4 days
   C) 7 days
   D) 10 days

   The circle next to “7 days” is filled in because there are 7 days in a week.

   For some questions you will be asked to write your answers in the space provided in your booklet. Example 2 shows this kind of question.

Example 2

2. Where does the little boy go after he finds the book?

   For questions like Example 2, you will see a 1 a 2 or a 3 next to it.
Example 3 shows a question with a \( \textcircled{3} \) next to it.

**Example 3**

3. What makes the ending of the story both happy and sad? Use what you have read in the story to help you explain.

\[ \textcircled{3} \]

You will have 40 minutes to work in your test booklet and then we will take a short break. Then, you will work for another 40 minutes. Do your best to answer all the questions. If you cannot answer a question, move on to the next one.
Stop

Do not turn the page until you are told to do so.
Questions  Hare Heralds the Earthquake

1. What was the hare's greatest worry?
   A) a lion
   B) a crash
   C) an earthquake
   D) a falling fruit

2. What made the whole earth shake?
   A) an earthquake
   B) an enormous fruit
   C) the fleeing hares
   D) a falling tree

3. Things happened quickly after the hare shouted “Earthquake!”
   Find and copy two words in the story that show this.
   1. ______________________________________
   2. ______________________________________
4. Where did the lion want the hare to take him?

5. Why did the lion drop the fruit onto the ground?
   - A to make the hare run away
   - B to help the hare get the fruit
   - C to show the hare what had happened
   - D to make the hare laugh

6. How did the hare feel after the lion dropped the fruit onto the ground?
   - A angry
   - B disappointed
   - C foolish
   - D worried
7. Write **two** ways in which the lion tried to make the hare feel better at the end of the story.

1. 

2. 

8. Do you think the lion liked the hare? What happens in the story that shows this?
9. How did the hare's feelings change during the story?

At the beginning of the story the hare felt

because

At the end of the story the hare felt

because
10. You learn what the lion and the hare are like from the things they do in the story. Describe how the lion and the hare are different from each other and what each does that shows this.

11. What is the main message of this story?

A) Run away from trouble.
B) Check the facts before panicking.
C) Even lions that seem kind cannot be trusted.
D) Hares are fast animals.

Stop

End of this part of the booklet.
Please stop working.
Questions  Nights of the Pufflings

1. Why are puffins clumsy at takeoffs and landings?
   - A  They live in a land of ice.
   - B  They hardly ever come to shore.
   - C  They spend time on high cliffs.
   - D  They have chunky bodies and short wings.

2. Where do the puffins spend the winter?
   - A  inside the cliffs
   - B  on the beach
   - C  at sea
   - D  on the ice

3. Why do the puffins come to the island?
   - A  to be rescued
   - B  to look for food
   - C  to lay eggs
   - D  to learn to fly
4. How does Halla know the pufflings are about to fly?

A. Parents bring fish to the pufflings.
B. Flowers are in full bloom.
C. Chicks are hidden away.
D. Summer has just begun.

5. What happens during the nights of the pufflings?

A. Puffin pairs tap-tap-tap their beaks together.
B. Pufflings take their first flight.
C. Puffin eggs hatch into chicks.
D. Pufflings come ashore from the sea.

6. What could the people in the village do to stop the pufflings from landing there by mistake?

A. Turn off the lights
B. Get the boxes ready
C. Keep the cats and dogs inside
D. Shine their torches in the sky
Questions 7 and 8 ask you to explain how Halla rescues the pufflings.

7. Explain how Halla uses her torch to rescue the pufflings.

8. Explain how Halla uses the cardboard boxes to rescue the pufflings.

9. According to the article, which of these is a danger faced by the pufflings?

A. drowning while landing in the sea
B. getting lost in the burrows
C. not having enough fish from their parents
D. being run over by cars and trucks
10. Why does it need to be daylight when the children release the pufflings? Use information from the article to explain.

11. What do the pufflings do after Halla and her friends release them?
   A. walk on the beach
   B. fly from the cliff
   C. hide in the village
   D. swim in the sea

12. Write two different feelings Halla might have after she has set the pufflings free. Explain why she might have each feeling.

   1. 

   2. 
13. Would you like to go and rescue pufflings with Halla and her friends? Use what you have read to help you explain.