EXAMINATION OF TEST EQUIVALENCE BETWEEN FRENCH AND ENGLISH LANGUAGE VERSIONS OF PROGRESS IN INTERNATIONAL READING LITERACY STUDY 2011

by

Shawna Goodrich

B.A., University of New Haven, 1988
M.S.W., Simmon's College, 1992

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

MASTER OF ARTS

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation and Research Methodology)

The University of British Columbia

(Vancouver)

December, 2013

ABSTRACT

In Canada, international large-scale assessments (LSAs), such as the Progress in International Reading Literacy Study (PIRLS), are administered in the two official languages, French and English. The validity of decisions made from these assessments depends on the equivalence of different language versions and the comparability of scores across language groups. Previous research examining French and English language versions of large-scale assessments administered in Canada indicates that equivalence cannot be assumed when tests are adapted (Ercikan, Gierl, McCreith, Puhan & Koh, 2004; Ercikan & McCreith, 2002; Gierl, 2000; Oliveri & Ercikan, 2011). Research has shown that the quality of test adaptation is particularly important to ensure comparability, interpretability and consequential equity across language groups.

The purpose of this study is to examine test equivalence and score comparability at the item and test level between French and English language groups administered PIRLS 2011. Confirmatory factor analysis and two methods of differential item functioning were conducted to examine score comparability. Four bilingual expert reviewers with expertise in reading literacy conducted independent blind linguistic and cultural reviews to identify the degree of test equivalence and potential sources of differences between the French and English language versions of released items from PIRLS 2011. As a whole, evidence from this study indicates there are important scale level differences between the French and English language versions of PIRLS 2011 that call for further investigation and on average 25% of items across thirteen booklets function differently at the item level. Reviews by experts of released items indicate that there are many differences between the two language versions for both statistically identified DIF items and non-statistically identified items. Reviewers concluded that inappropriate

translation produced unintended differences in content and difficulty levels between the two language versions.

PREFACE

This dissertation is the original, unpublished, independent work of the author, Shawna Goodrich.

The work contained in this thesis was approved by:

UBC Research Ethics Board: Behavioural Research Ethics Board

Project Title: Examination of Test Equivalence Between French and English Language Versions of the Progress in International Reading Literacy Study 2011

Certificate Number: H13-01719

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ACRYNOMS

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
CMEC	Council of Ministers of Education, Canada
CFA	Confirmatory factor analysis
CR	Constructed response
DIF	Differential item functioning
DTF	Differential test functioning
EFA	Exploratory factor analysis
ICC	Item characteristic curve
IEA	International Association for the Evaluation of Educational
IRF	Item response function
IRT	Item response theory
ITC	International Test Commission
LR	Logistic regression
LSA	Large scale assessment
MC	Multiple choice
OECD	Organization for Economic Co-operation and Development

OLR	Ordinal Logistic Regression
PIRLS	Progress International Reading Literacy Study
PISA	Programme for International Student Assessment
RMSEA	Root mean square error of approximations
RMSR	Root mean square residual
TTC	Test characteristic curve
TIMSS	Third International Mathematics and Science Study

ACKNOWLEDGMENTS

Nothing worthwhile is accomplished alone. I have many wonderful people to thank for helping me to complete this thesis. Foremost, is my husband, Vikram Dua, who worked as hard as I did attending to our children, animals, home and my emotional well being. Together, as with all aspirations, we accomplished this work. I would also like to thank my children Hannah and Akshay who sacrificed a great deal to accommodate my unrelenting schedule and who constantly expressed their conviction that their mother could accomplish anything she set her mind too. Your belief sustained me. I would like to thank my mother for instilling in me a love of reading literacy that has infused my life and this thesis.

Next I would like to thank my supervisor, Dr. Kadriye Ercikan, for showing me the way before I knew where I was going and for providing mentorship and guidance to persevere in the face of obstacles. I would also like to thank my committee members, Dr. Nancy Perry and Dr. Bournot-Trites for their insightful feedback and willingness to work within such a tight timeframe.

I also would like to thank the women in my life who were and are always by my side cheering me on and bolstering me. I have Mary to thank for her impertinence and incessant amusement. I have Juliette to thank for her sensible wisdom and companionship. Many other friends including Lisa, Sarah, Fawn and Dallie also deserve my gratitude for their continual support and encouragement. And, as always, I would like to thank my best friend, Josie, who supports me through it all.

Chapter I

International large-scale assessments (LSAs), such as the Progress in International Reading Literacy Study (PIRLS), are administered in 49 countries across 48 languages for the purposes of evaluating educational systems, informing curricular planning, resource allocation and setting educational policy and practices. Results from LSAs serve two general purposes. Within countries results are used to draw conclusions about academic achievement, learning and educational accountability (Crundwell, 2005). Across countries LSAs are used to draw comparative conclusions about academic achievement, learning and educational accountability. Large-scale assessments provide evidence to policymakers and administrators about the strengths and weaknesses of educational systems. Evidence from LSAs informs decisions about the structure and delivery of education within nations and acts as initiators of reforms for educational programs. In consideration of the influential role of international LSAs on educational policy worldwide, it is essential to ensure that tests are equivalent across linguistic and cultural groups and scores are comparable. When tests are not equivalent and scores are not comparable across languages, decisions and consequences that are based on results may be unfair, unjustified and misguided (Bond, Moss & Carr, 1996). The *Standards for Educational and Psychological Testing* define large-scale assessments as "any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects or programs" (AERA, APA, & NCME, 1999, p. 172).

The increased use of multiple language versions of assessments has given rise to methods, standards and guidelines to aid in the development and administration of adapted tests (Hambleton, 2005). These include the *Standards for Educational and Psychological Testing* (2012), referred to as the *Standards* hereafter, developed by the American Educational Research

Association (AERA), the American Psychological Association (APA) and the National Council for Measurement in Education (NCME) and the *International Test Commission Guidelines for Test Adaptation* (ITC, 2000), referred to as the *Guidelines* hereafter. Both provide careful and detailed instructions about how to adapt tests from one dialect, language and culture to another (Hambleton, 2005). They were developed to ensure that appropriate procedures are utilized to verify test equivalence and score comparability across languages (Ercikan, Simon & Oliveri, 2013; Sireci, Yang, Harter & Ehrlich, 2006). However, despite the widespread use of multiple language versions of LSAs and the increased attention given to test adaptation and measurement comparability, research indicates that equivalence cannot be assumed when tests are adapted (Allalouf, Hambleton & Sireci, 1999; Arim & Ercikan, 2005; Ercikan & McCreith, 2002; Ercikan et al., 2013; Oliveri & Ercikan, 2011; Seller, Gafni & Hanani, 2005). Research conducted in Canada comparing the French and English versions of international LSAs have recurrently found that 18 to 60% of the items function differently for the two groups (Ercikan, Gierl, McCreith, Puhan & Koh, 2004; Ercikan & McCreith, 2002; Gierl, 2000; Oliveri & Ercikan, 2011; Rogers & Klinger, 1999). Such research indicates that French and English language test versions are not equivalent.

Given the complexities of test adaptation, the *Standards* outline four recommendations to ensure measurement equivalence across linguistic and cultural groups (Ercikan & Lyons-Thomas, 2013; Geisinger, 1994; Hambleton, 2005). In recent years, the term test translation has been replaced with test adaptation in reference to the breadth of activities that are necessary to create different language and cultural versions of a test. *Standard 9.4* recommends that if linguistic modifications are made on a test, publishers provide justification for modifications and address implications to score interpretations. *Standard 9.5* advises test developers to provide

additional information to assist test users in the interpretation of test scores, if there is evidence that scores are not comparable across test versions. *Standard 9.7* recommends that test developers describe the procedures used to establish and ensure adequacy of adaptation and provide evidence of score reliability and validity for linguistic groups. *Standard 9.9* recommends that if test developers intend to create comparable linguistic versions of a test, evidence of test comparability should be provided.

These standards provide a framework for considering the types of error that can occur when a test developed for one particular group is adapted for a different linguistic and cultural group. When tests are adapted, errors are inevitable but they may not necessarily bias test results (Arffman, 2013). The concept of bias refers to "the extent to which there is evidence of differential validity for any relevant subgroup of persons affected (Bond, Moss & Carr, 1996, p. 17)." There are three forms of bias that can jeopardize the validity of test scores. They include construct, method and sample bias. Of the three, the most fundamental form of bias to group comparisons is construct bias. It is essential to ensure that a test is measuring the same ability, trait or construct consistently for all groups of interest. Construct bias implies differences in the essential elements of constructs across linguistic and cultural groups (van de Vijver & Poortinga, 2005). These essential differences may include culturally specific norms, behaviors and attitudes. For example, a researcher interested in comparing stress responses to unemployment in a sample of Canadians and Koreans may find culture specific behavior differences associated with stress responses. Method bias is another form of bias that results from differential familiarity with administrative test procedures between groups. There are a number of types of method bias that include sample, instrument and administrative bias (van de Vijver & Poortinga, 2005). Sample bias can occur when two linguistic groups are tested for reading literacy but one group is more

fluent in the language of the test, such as is often the case in the assessment of English language learners. Item bias is the third form of bias and occurs when a characteristic of the item is not relevant to the underlying ability or construct that is being measured and is unfair to one or more groups (Sireci & Allalouf, 2003). For instance, if the vocabulary of an item is more difficult in one language version than the other language version then the item may be biased.

Research has shown that the quality of the adaptation process directly impacts the validity of measurement and score comparability (Arffman, 2010; Cook, Schmitt-Cascallar & Brown, 2005; Sireci, Yang, Harter & Ehrlich, 2006; Solano-Flores, Trumbull & Nelson-Barber, 2002; van de Vijver & Tanzer, 2004). Score comparability refers to the degree to which scores from source and target versions of tests have the same meaning or demonstrate the same level of performance proficiency on content standards (Bowles & Stansfield, 2008). In a bilingual country, such as Canada, where tests are administered in the two official languages, French and English, the quality of test adaptation is particularly important to ensure comparability, interpretability and consequential equity for both language groups. It is incumbent on countries with two official languages to take reasonable steps to ensure that linguistic groups are given the opportunity to perceive and respond to tests in the same way (Fairburn & Fox, 2009; Rogers, Lin & Rinaldi, 2010). Although research shows that the quality of the adaptation process is instrumental to establish equivalence across linguistic test versions, there are numerous factors that contribute to nonequivalence.

Factors Associated with Nonequivalence

Non-equivalence across test versions may be attributable to a range of factors that include cultural and curricular differences; however, language differences create many of the challenges associated with test adaptation (Bachman, 2000; Cohen, 2007; Cummins, 2000; Solano-Flores,

Contreras-Niño & Backhoff, 2013; Tanzer, 2005). This is due in part to the inherent differences between languages. Languages differ in both form and meaning (Steiner & Mahin, 1996). Form relates to sentence structure, writing systems, word order, ways of conveying new information, ways of signaling thematic structures and methods of cohesion (Arffman, 2007; Baker, 1992). Meaning is inextricably interrelated with language (Rorty, 1977). Human beings acquire and use language to explore and create meaning by interpreting experiences and engaging in interpersonal relationships (Halliday, 1993). Likewise, meaning is created through language. Every language system attaches different meanings to different aspects of language including grammar, syntax and semantics. The meanings that are attached to words are a function of a language as a whole and language is interwoven with social and cultural practices (Roth, Oliveri, Sandilands, Lyons-Thomas & Ercikan, 2013). Previous research has identified a number of differences between languages that make test adaptation difficult and can result in incomparability of assessment scores between language groups (Sireci, 2008). These include differences in grammar, meaning, vocabulary, syntax, word usage and difficulty (Allalouf et al., 1998; Allalouf & Hambleton, 1999; Arffman, 2007; Ercikan, 1998; Ercikan et al., 2004). Research has shown that language specific grammar differences can result in the nonequivalence of item difficulty between different language versions of tests (Arffman, 2007; Bonnet, 2002; Ercikan, 1998; Grisay & Monseur, 2007).

Research has also demonstrated that psychometric differences between language versions may be attributable to factors other than test adaptation, such as cultural and curriculum differences between groups (Ercikan et al., 2004; Solano-Flores & Nelson-Barber, 2000 & 2001). Examinees not only draw upon their language to make sense of words and texts but their social and cultural experiences (Arffman, 2007; Gee, 2001; Greenfield, 1997; Steiner & Mahn,

1996). Words are not inherently meaningful; social and cultural interactions and conventions imbue words with meaning (Campbell, 2003; Derrida, 1998; Greenfield, 1997). Cultural experiences may produce different interpretations of commonly shared words, which affect the trajectory of thought processes and ultimately responses to test questions (Ercikan & Lyons-Thomas, 2013; Roth, 2009; Solano-Flores, 2006). Such research has drawn attention to the importance of understanding what causes psychometric differences between language versions of tests. By identifying the source of differences, it may be possible to reduce the number of items that function differently across language groups (Bolt & Gierl, 2006; Ercikan, Arim, Law, Domene, Lacroix, 2010; Ercikan & Lyons-Thomas, 2013; Gierl & Khaliq, 2001). There are a number of quantitative and qualitative methods to examine score comparability. In the next section, the methods used in this study are discussed briefly.

Methods to Examine Score Comparability

As previously mentioned, fundamental to score comparability and the overall objectives of LSAs is the meaningful interpretation of test scores. To make comparative inferences across languages and cultures on the basis of test scores requires evidence of the comparability of test scores (Gierl, 2000; Hambleton, 2005; Sireci, 2005; Sireci, Patsula & Hambleton, 2005). Evidence must demonstrate that tests are measuring the same construct and they are doing so adequately at a comparable level of difficulty (Arffman, 2007; Sireci, 2005; van de Vijver & Poortinga, 2005). Evidence must also demonstrate that test scores from different language versions are on a common scale.

Statistical strategies, in conjunction with expert reviews, are the most common methods that are used to examine score comparability and to identify and interpret sources of bias (Benítez & Padilla, 2013; Ercikan et al., 2004; Ercikan et al., 2010; Sireci et al., 2005 & 2006).

Statistical analysis can be conducted to examine construct, method and item bias. For example, at the test level, measurement equivalence can be examined through the use of dimensionality analysis (Gierl, 2005; Koh & Zumbo, 2008). Conducting exploratory factor analysis (EFA) or confirmatory factor analysis (CFA) provides evidence to examine the similarity of factor structures across the two groups. If items load equivalently on the factors that account for the variation of the test data, this provides evidence to support similarity of construct measurement for the groups. Items that load differently across groups indicate factor structure variance, which may suggest that construct measurement is different for the two language groups or that the factor model is mis-specified in one or both groups (Meade & Kroustalis, 2006: Zumbo, 2003).

Differential item functioning (DIF) analysis is a common statistical technique used to examine the comparability of measurement for different language groups at the item level. If the probability of responding correctly to a single item differs for groups with equal proficiency levels, the differences may reflect some characteristic of the item rather than ability differences. Differences in item responses are investigated after groups are statistically matched on the characteristic of interest. Again, although test score differences between linguistic groups may reflect real discrepancies in knowledge and competencies, DIF captures psychometric differences in test items for matched ability groups. Research has identified large amounts of DIF across different language versions of LSAs (Ercikan, 2002; Gierl, 2000; Oliveri & Ercikan, 2011).

Another recommended method to examine comparability of test items is bilingual expert reviews of items. Expert bilingual reviews are used to identify potential sources of DIF and to confirm differences identified by statistical analyses (Ercikan et al., 2004; Ercikan et al., 2013; Gierl & Khaliq, 2001; Puhan & Gierl, 2006). Expert reviews are a crucial component of

measurement comparability research that can be used to facilitate meaningful interpretations of score differences and to create comparable linguistic test versions (Ercikan et al., 2010; Puhan & Gierl, 2006).

Significance and Purpose of the Study and Research Questions

In consideration of the increased influence that LSAs have on curriculum, educational policy and resource allocation in Canada it is imperative that tests are equivalent across French and English language versions. If tests are not equivalent across language versions in Canada, inferences are not valid and scores are not comparable. As the official languages in Canada are French and English, the implications of test inequivalence to Canadian society are significant. The implications include invalid inferences based on language comparisons that lead to inappropriate decisions about resource allocation, curriculum planning and educational policy. When LSAs contain linguistic and cultural differences that unfairly favor one group over another, inferences based on score comparability are invalid. In consequence, decisions regarding educational policy, curriculum and resource allocation that are based on invalid inferences are likely to be ineffective and inefficient (Padilla & Medina, 1996). Invalid inferences can also have direct and indirect negative effects on students' academic attitudes, achievements and motivations (Padilla & Medina, 1996). Given the potential implications of test inequivalence across French and English languages in Canada, the focus of this study is on the only international LSA program to assess reading literacy in early years of schooling in Canada. The Progress International Reading Literacy Study (PIRLS) is administered in Canada to students in 4^{th} grade to determine achievement levels of students at the classroom, school, regional, provincial and national levels. Scores from PIRLS are used to distinguish differences between identifiable populations.

The purpose of this study is to examine test equivalence of the French and English language versions of PIRLS 2011 administered in Canada to evaluate the extent of score comparability between these two groups. To examine the equivalence of French and English language versions of PIRLS 2011 statistical analysis and expert reviews were conducted. Both CFA and DIF IRT were conducted to examine the equivalence of the construct assessed by PIRLS at the scale and item level. Research has indicated that the inclusion of CFA and DIF IRT results in a comprehensive construct equivalence examination (Allalouf et al., 1998; Gierl et al., 1999; Zumbo, 2008). Research also indicates discrepancies between item and test level measurement comparability (Oliveri, Olson, Ercikan & Zumbo, 2012; Zumbo, 2003). Two DIF detection methods were used to examine items from the French and English language versions of PIRLS 2011. Finally, expert reviews were conducted for items and passages publicly released from the administration of PIRLS 2011. As only four of ten reading passages from PIRLS 2011 have been released to the public, expert reviews were restricted to these four reading passages and items accompanying them. Without access to all ten passages it is not possible to identify all of the DIF sources in this study, so expert reviews are used to examine the potential sources of DIF for the released items from the four passages. A detailed description of the passages and the assessment framework used for PIRLS 2011 can be found in chapter 3, the method's section.

The research questions of this study follow:

1. To what extent is the factor structure of the construct evaluated by PIRLS 2011 equivalent across the French and English language groups?
2. To what extent do items function differently across the French and English language versions of PIRLS 2011 administered in Canada?

3. What do expert reviews reveal about the equivalence of French and English released items from PIRLS 2011 administered in Canada?

In the next chapter theoretical and empirical research on test equivalence and score comparability are discussed. A brief literature review of recommended methodologies to investigate test equivalence is also discussed. The chapter ends with an overview of reading literacy in a sociocultural context.

Chapter II: Literature Review

The focus of this chapter is on problems associated with the adaptation of LSAs and the implications of such problems on test equivalence and score comparability, with a specific emphasis on the PIRLS for French and English language groups in Canada. This includes a discussion and review of literature related to concepts that are fundamental to the current study. The chapter begins with an overview of LSAs before the specific uses and purposes of LSAs in Canada are discussed. Definitions of score comparability and test equivalence as conceptualized in test theory and a review of empirical research examining the equivalence of LSAs follows. Then methods for examining item and test equivalence are reviewed briefly and in detail in the next chapter. The concept of test equivalence is developed from the perspective of test theory and test translation theory (Arffman, 2010; Solano-Flores, Backhoff & Contreras-Nino, 2009). This chapter concludes with a discussion of reading literacy in the context of social and cultural factors that shape the meaning of language and reading.

Large-Scale Academic Achievement Assessments

A firmly held belief among policy makers is that there is a direct relationship between the merits of an education system with respect to assessment results and how successful a country is from an economic standpoint (Bonnet, 2002). Interest in international large-scale assessments (LSAs) as comparative indicators for gaging educational results is due in part to this belief. The Organization for Economic Co-operation and Development (OECD) stresses the value of comparative indicators as tools to formulate national education policies (Bonnet, 2002). Increasingly, the results of international LSAs are the basis from which nations make educational decisions regarding instruction, curriculum and policy decisions (Arffman, 2007; Bonnet, 2002;

Shohamy, 2007; Stobart, 2003). The number of countries participating in international LSAs, such as PIRLS, has increased steadily in the last decade.

The increasing demand for accountability in education has lead to the rise of reading literacy LSAs (Bachman, 2007). Most LSAs are created to have different but comparable forms for use across multiple language and cultural groups. For reading literacy assessments to be administered across diverse linguistic groups requires translation from the original *source* version in English to a *target* language version. For the results to be valid and the decisions based on them to be fair and beneficial for all linguistic groups, it is essential to ensure that different language versions of a test measure the intended construct and are equivalent to each other. To date, research shows that test translation and adaptation does not ensure that linguistic versions are equivalent across tests (Ercikan, 2002; Ercikan et al., 2004; Ercikan & Koh, 2005; Gierl & Khaliq, 2001; Maldonado & Geisinger, 2005; Ockey, 2007; Yildirim & Berberoglu, 2009). Research also suggests that it may be impossible to create adapted LSAs that are free from linguistic and cultural bias, specifically for reading literacy tests (Arffman, 2010; Bonnet, 2002; Solano-Flores & Allalouf, 2003).

It is important to note that the consequences of results for LSAs vary. For instance, low-stakes tests are often administered to assess curriculum and/or grade level expectations at regional, national and international levels, as well as to make comparative interpretations across countries, and they do not have consequences for individual examinees, teachers and schools. Conversely, results from high-stakes tests are used to make decisions about examinees with clear consequences (Haladyna & Downing, 2004). The need to establish score comparability increases as consequences based on inferences from test scores increase (Madaus & Clarke 2001).

Large-Scale Academic Achievement Assessments in Canada

Across the 10 provinces and two of the northern territories of Canada, the number of LSAs (provincial/territorial, national and international) administered to students throughout their elementary and secondary education has increased significantly (Ercikan, Oliveri & Sandilands, 2012; Klinger, DeLuca & Miller, 2008; Volante & Jaafar, 2008). In addition to administrating international LSAs, every province and territory in Canada administers at least one LSA. Accompanying this increase is a growing influence of LSAs on all aspects of Canadian education including curriculum, instruction, school accountability, performance standards and educational policy (Klinger et al., 2008). Each province and territory is responsible for the development of curriculum and the assessment of student achievement. Unlike in the U.S., where there is a governing body that unifies educational policies, the specific structure and general organization of the educational system varies across the provinces/territories of Canada (Volante et al., 2008). The decentralized state of education in Canada is thought to reflect the cultural, historical and geographical diversity across the country. However, the linguistic, cultural and curricular differences across Canadian provinces/territories create challenges for within country test score comparisons (Klinger et al., 2008). Furthermore, the governance structure of education across Canada poses challenges for balancing the demands of international LSA trends that emphasize centralization and standardized accountability (Volante et al., 2008).

In Canada, concerns are being raised regarding the uses and limits of LSAs (Ungerleider, 2006). Consequential validity of testing or the increased use of LSAs for purposes that extend beyond original intentions is a particular concern with respect to the appropriateness of interpretations and actions based on test scores (Stobart, 2003). Messick (1995) articulates a unified framework in which he argues validity is an evaluative judgment of the degree to which

evidence can support the uses, interpretations and consequences of test scores. Score meaning is a construction that must make theoretical and empirical sense. Inferences about what test scores mean must be validated. Actions and consequences that are based on the meanings of test scores must be justified by evidence. When score-based interpretations are extrapolated beyond the original test context, validity for alternative uses must be provided to support decision-making.

In Canada, PIRLS is the only international LSA administered to assess reading literacy in the early years of education. The Council of Ministers of Education (CMEC, 2012) specifies that score results from PIRLS inform educational research and policy in Canada. More specifically, results from PIRLS are used to track early literacy skills of identifiable subpopulations over time; evaluate changes implemented in educational systems, and provide information to identify and remedy structures and processes that limit and enhance reading achievement across levels of educational systems (Labrecque, Chuy, Brochu & Houme, 2012). In the PIRLS 2011 cycle, three of the nine provinces, Alberta, Ontario and Quebec, participated at the 'benchmarking' level (Labrecque et al., 2012), which enables provinces to compare and evaluate achievement in an international context. In addition, students from four provinces, British Columbia, New Brunswick (French), Nova Scotia and Newfoundland and Labrador, were oversampled in 2011 to allow for analysis by subgroups. Oversampling provides more reliable estimates to allow for comparative analyses of each subgroup within Canada and internationally (Labrecque et al., 2012).

PIRLS data is unique because it enables within country comparisons of grade 4 reading achievement and tracking of literacy levels over time. The uses and benefits of PIRLS data are outlined in the 2012 CMEC report:

PIRLS allows Canadian jurisdictions not only to evaluate the changes implemented in their educational systems but to also consider them in an international context. Results obtained by PIRLS should help channel spending to those areas of early education where it is most needed (Labrecque et al., 2012, p. 5).

The CMEC report also states that participation allows for the collection of reading proficiency level data to enable educators to intervene earlier. As an illustration of how PIRLS data can help provide information to relevant stakeholders, the report makes note of the decreased reading proficiency level of grade 8 students enrolled in French schools (Labrecque et al., 2012). French language Canadian students performed more poorly compared to English language Canadian students according to a number of studies on international LSAs (Bussiere, Cartwright, Knighton & Rogers, 2004; Ercikan et al., 2004; Ercikan & McCreith, 2002; Gierl, 2000). Given the performance differences between French and English language groups in Canada and the stated uses of PIRLS data, it is essential to ensure that conclusions and decisions based on comparative score results are valid. Test equivalence between French and English linguistic versions and score comparability are central to the validity of PIRLS results.

Overall, Canadian students performed well on PIRLS 2011, with an average achievement score of 548 SE 1.6, based on an international mean of 500 and SD of 100 (Labrecque et al., 2012). Only seven countries obtained averages higher than Canada, with the highest scores from students in Hong Kong (571) SE 2.3 and the second highest from students in the Russian Federation (568) SE 2.7. Some of the lowest average scores are from French speaking countries such as France 520 SE 2.6 and French Belgium 506 SE 2.9. In relation to the Canadian average only British Columbia performed above the Canadian average, while Ontario, Nova Scotia, Alberta, Newfoundland and Labrador performed at the Canadian average. The average scores for

Quebec (538) SE 2.1 and French New Brunswick (514) SE 2.7 were significantly lower than scores of Canada overall. To examine differences between language groups four provinces oversampled students from French and English-language school systems. The only provinces to score significantly below the Canadian average were French Quebec and French New Brunswick. The average score differences were all in favor of English language students.

In British Columbia, students from English-language schools scored an average of 43 points higher or approximately 0.5 SD difference (SD = 100) than students from French-language schools, in Ontario the average difference between English and French language groups was 48 points, in Quebec the average difference was 8 points and in Nova Scotia 51 points. English speaking students in Quebec performed slightly more poorly than English speaking students in other areas of the country, as shown in Table 1.1. The table below presents the average PIRLS 2011 scores and differences for students enrolled in French and English language school systems.

Table 2.1

Average Scores and Differences by Language

	English language school system		French language school system		
	Average score	SE	Average score	SE	Difference
BC	556	3.2	513	6.2	43
AB	548	2.8	_	_	_
ON	554	2.7	506	3.5	48
QC	545	3.6	537	2.4	8
NBf	_	_	514	2.7	_

	English language school system		French language school system		
	Average score	SE	Average score	SE	Difference
NS	551	2.5	500	3.7	51
NL	547	2.8	_	_	_
CAN	553	2.0	533	2.1	20

Note. SE = standard error. This table is from PIRLS 2011 Canada in Context (CMEC, 2012)

It is useful to compare score differences from PIRLS 2011 to score differences from the Programme for International Student Assessment (PISA) 2009. Like PIRLS, PISA is an international LSA administered in 65 countries to provide comparative information. However, PISA assesses one major and two minor subject domains every three years. In 2009, reading was the major domain assessed by PISA. An important difference between PIRLS and PISA is that the sampled age population for the former is 9 or 10 years old and for the later it is 15 years old. The type of text used in the two assessments also differs but the reading processes that were measured are similar. As with PIRLS, Canadians performed well in reading on PISA 2009 ranking as the 6th highest country, with an average achievement score of 524 (Knighton, Brochu, & Gluszynski, 2010). In the provinces of Ontario and Alberta, Canadians performed above the Canadian average, while British Columbia and Quebec performed at the Canadian average. Reading performance was compared in seven provinces for students in French and English-language school systems (British Columbia, Alberta, Manitoba, Ontario, Quebec, New Brunswick and Nova Scotia). In contrast to PIRLS, there were no significant score differences between language groups in Quebec and Manitoba. Similar to PIRLS, there were significant score differences in favor of students in English-language school systems in five provinces. The average score differences between language groups for PIRLS and PISA were consistent in four

of these provinces. Overall a few of the score patterns for French and English language groups were similar for PIRLS and PISA, but there were also important differences in Nova Scotia, Quebec and Manitoba.

Score Comparability and Equivalence

International LSAs are criticized for a number of reasons. They have been criticized as being biased in favor of Western and Anglo-Saxon culture because they are funded by western organizations, modeled by western dominated psychometric views and developed in English (Goldstein & Thomas, 2008; Murat & Rocher, 2004; Solano-Flores, 1999, 2000, 2011; Tanzer, 2005; Van de Vijver and Leung, 1997). Some cross-cultural assessment researchers argue that current paradigms limit the possibility of obtaining accurate information on examinees outside of the dominant culture (Solano-Flores et al., 2001). In the *Standards,* bias is conceptualized as group differences that result from deficiencies in a test or the way that a test is used, rather than on true ability differences between groups. Bias occurs when test items contain sources of difficulty that are irrelevant to the construct that is being measured (Zumbo, 1999). When performance differences are due to bias, score comparability may be jeopardized.

The *Standards* were created to provide criteria to evaluate the development of tests, testing practices and the effects of test use (AERA, APA & NCME, 1999). The *Guidelines* (ITC, 2001) were created to aid in the adaptation of tests for use with different linguistic and cultural groups. Both the *Standards* and the *Guidelines* emphasize that the *quality* of test adaptation affects the validity of comparative inferences (Hambleton, 2005; Sireci et al., 2006). As previously mentioned, other factors in addition to test adaptation result in score incomparability across linguistic groups. These may include interest in and familiarity with the content of items

due to cultural differences and curriculum differences that result in varying degrees of exposure to test content between groups.

Previous research has shown that test adaptation results in significant score incomparability (Ercikan, 1998; Ercikan, 2002; Ercikan et al., 2004; Ercikan et al., 2010; Gierl & Khaliq, 2001; Maldonado et al., 2005; Yildirim et al., 2009). For this reason the *Guidelines* and the *Standards* require that equivalence and fairness be established when an instrument is adapted from one language to another (Cook, Schmitt-Cascallar & Brown, 2005). Adaptation of a test does not ensure equivalence between target and source versions (Beller et al., 2005; Cohen, Gafni & Hanani, 2007; Hambleton, 2005; Solano-Flores, 2006). In fact, test adaptation can produce unintended differences in content and difficulty levels between linguistic versions of a test, which contribute to observed score differences. For instance, in a study that examined the comparability of the French and English versions of PISA 2000 in Canada reported significant differences in the word and character count between the two versions, which had a moderate effect on the difficulty of the items for the French version (Grisay, 2003).

Different levels of equivalence in test theory determine the measurement level at which scores can be compared and the types of comparative inferences that can be made across language groups (Sireci, 2005; van de Vijver et al., 2005). Equivalence in test theory refers to the degree to which test scores can be used to make comparable inferences for different examinees (AERA, APA & NCME, 1999). In this study, test equivalence of the English and French versions of the PIRLS 2011 is examined at item and test levels. Test equivalence indicates the extent to which the ability, trait or construct being measured is the same for the groups being compared (Gierl, 2000; Hambleton, 2005). The validity of all comparative conclusions based on international LSAs is dependent on the assumption that test versions measure the same construct

with the same level of measurement accuracy. To establish equivalence of the concepts and definitions across groups requires both theoretical and empirical evidence (Hambleton, 2005). Measurement incomparability across linguistic groups for international LSAs is not uncommon (Ercikan & Koh, 2005; Gierl, 2000; Sireci et al., 2003). In fact, measurement differences are not surprising when groups from different countries with different cultures and curriculums are compared; however, research also shows differences among linguistic groups within a single country. In a study that examined measurement comparability for the English and French language versions of the Third International Mathematics and Science Study (TIMSS), differences were found in the constructs measured in Canada (Ercikan et al., 2005). Differences between French and English Canadian groups resulted in the detection of DIF for 14% of the mathematic items and 37% of the science items, with 0.40 correlations between discrimination parameters of the two language groups for science items.

Test equivalence between source and target versions of a test incorporates linguistic, cultural and content equivalence. To establish test equivalence requires evidence that a test measures the same construct, that the construct has similar meanings across groups and the test format and length are the same across groups (Ercikan & Lyons-Thomas, 2013). To establish test equivalence requires the use of statistical analysis to identify items that function differently across languages and qualitative analyses to identify the potential causes for the differences (Lazaraton & Taylor 2007). In a study that examined test equivalence between a Hebrew source version and a Russian target version of a high stakes university admission test administered in Israel statistical analyses identified 42 out of 125 items as DIF (Sireci et al., 2003). Potential causes for item differences were examined through follow-up qualitative analysis using five Hebrew-Russian expert reviewers. Of the 42 items identified as DIF, expert reviewers identified

changes in word difficulty resulting from test adaptation as a potential cause for 16 DIF items. The cause for another eight DIF items was identified as potentially resulting from changes in the content or meaning of items associated with adaptation. Sources for the remaining DIF items were judged to result from changes in format and differences in cultural relevance. Such studies emphasize the importance of identifying potential sources once DIF is detected by statistical methods.

Testing conditions equivalence is another form of equivalence that requires: a) tests are administered in the same manner across groups; b) the test format be equally appropriate across groups; c) the speed of responses not have a greater effect for one language version; and d) response styles do not differently affect groups. Evidence for the administration of tests requires data and documentation of the procedures used. Evidence for format equivalence should be documented during the test development phase but can also be evaluated through qualitative analysis after administration. The speed of responses can be examined during test examination or by conducting psychometric analyses of test response time.

Methods for Examining Score Comparability

One of the primary statistical methods used to examine item bias across linguistic versions of assessments is differential item functioning (DIF) analysis. DIF methods are used to address the question of whether items function differently across groups of examinees with equal proficiency levels. Differences in item responses are investigated after groups are statistically matched on the characteristic of interest (Linn & Harnisch, 1981). Although, test score differences between linguistic groups may reflect real discrepancies in knowledge and competencies, research has identified large amounts of DIF across different language versions of LSAs (Ercikan, 2002, 2003; Oliveri et al., 2011). Research demonstrates that differences across

groups are often attributable to a lack of equivalence across language versions (Ercikan et al., 2013; Oliveri et al., 2011). For instance, research conducted in Canada comparing the English and French versions of the PISA Problem-Solving Measure (PSM) have recurrently found DIF ranging from 18 to 60% (Ercikan, 2002, 2003; Ercikan et al., 2004; Gierl et al., 1999; Oliveri et al., 2011).

Another important finding in relation to DIF analyses is that the identification of items can vary with detection methods (Ercikan, 1999; Ercikan et al., 2004; Gierl et al., 1999; Oliveri et al., 2011). A number of test characteristics can affect the reliability of DIF statistics including the range of item difficulties, the distribution of abilities, sample size and differences in procedures for the estimation of DIF and the extent of DIF (Hambleton, 2006; Yildirim et al., 2009). Given the evidence of inconsistencies between DIF methods, the use of more than one method is recommended to ensure simultaneous detection across methods (Ercikan et al., 2004; Hambleton, 2006; Oliveri et al., 2011).

Research also indicates discrepancies between item and test level measurement comparability (Ercikan et al., 2005; Oliveri et al., 2012; Zumbo, 2003). Zumbo (2003) recommends the use of confirmatory factor analysis (CFA) and DIF analysis to ensure the comparability of construct at the scale and item level. In a study that examined the degree of construct comparability between the item and test level for two language versions of PISA, results suggested that different conclusions might be drawn from different levels of measurement comparability analysis (Ercikan et al., 2005; Oliveri et al., 2012). Discrepancies may be related to the cancellation of DIF with some items favoring one language group and other items favoring the other language group. For that reason, the use of CFA and DIF are recommended to obtain comprehensive construct equivalence information (Gierl et al., 1999; Sierci et al., 1998).

Research has also demonstrated that psychometric differences between language versions may be attributable to multiple factors (Arffman, 2010; Elosua & López-Jaúregui, 2007; Ercikan et al., 2002; Ercikan et al., 2004; Sireci et al., 2005; Solano-Flores et al., 2009; Wu & Ercikan, 2006). Such research points to the importance of using qualitative methods to examine the sources of DIF (Bolt et al., 2006; Ercikan et al., 2002; Ercikan et al., 2010, Gierl et al., 2001). The use of expert reviews to identify sources of DIF is essential to provide meaningful interpretations of score differences and to create comparable linguistic versions of assessments (Ercikan et al., 2010; Puhan et al., 2006). A number of qualitative methods are used to investigate sources of DIF and to confirm statistical analysis; however, expert reviews are the most commonly used method (Ercikan et al., 2004; Ercikan et al., 2013; Gierl et al., 2001; Puhan et al., 2006). Reviewers are instructed to conduct item reviews and categorize adaptation errors and rate equivalence according to a set of criteria. Items may be evaluated for content, equivalence of meaning, difficulty of vocabulary and sentence structure, differences in the length of words and sentences, and differences in item format and visual layout and omissions or additions that change the meaning or guide student thinking (Campbell & Hale, 2003; Ercikan et al., 2013; Puhan et al., 2006). However, research on DIF items using statistical and judgmental reviews has shown that reviews give inconsistent results. Some argue that this is due in part to a lack of clear standards or instructions to conduct expert reviews (Arffman, 2012; Campbell et al., 2003). For example, Grisay et al., (2007) found that review methods used in the development of PISA 2000 and PISA 2001 did not correctly predict DIF. Comparing item difficulties and the magnitude of DIF between national and international item parameters from PISA 2000 and 2001 studies they found a significant item-by-language interaction within language groups. For instance, certain items showed a significant item-by-language interaction for all the French

versions, which was not detected by the expert review process. Next is a discussion of emerging research on test translation error and expert reviews.

Theories of Test Equivalence

Test translation error. In view of the evidence that test adaptation creates potential sources of bias, systematic approaches to ensure proper implementation of guidelines are central to examine test equivalence across languages. Guidelines for adaptation appear to be insufficient to ensure high-quality adaptation (Arffman, 2010; Solano-Flores, Backhoff & Contreras-Nino, 2009). Members of the measurement community have recently been developing theoretical frameworks and analytical tools to guide in the adaptation and review of tests (Arffman, 2010; Solano-Flores et al., 2009).

Solano-Flores, Contreras-Niño and Backhoff (2005) developed a theoretical framework referred to as test translation error (TTE). The TTE framework draws on a broad scope of knowledge bases including the *Standards* and the *Guidelines,* the American Translator Association and academic fields such as linguistics, psychometrics, educational curriculum and cross-cultural assessment. In the TTE framework, equivalence between linguistic versions incorporates a variety of test item properties that may include content, language, format, conventions and linguistic demands. The TTE framework also accounts for other factors that are beyond the control of translators, such as tight deadlines that prevent translators from refining and reviewing the wording of items. Within this framework, there is an expectation that translation errors are unavoidable because languages encode meaning in different ways (Campbell & Hale, 2003). Minor translation errors such as slight variations in format may not necessarily produce measurement error (Solano-Flores et al., 2009). This framework and related research has fostered debate about what constitutes an acceptable level of translation error given

the content and purpose of a particular test.

The TTE framework is based on five principles: 1) inevitability of test translation error; 2) translation error dimensions; 3) relative dimensions of mistakes; 4) multidimensionality of translation errors; and 5) acceptability of translated item based on probabilistic nature. Error dimensions consist of broad categories of errors that test reviewers construct to comply with criteria. A specific set of dimensions that reflect the goals and content of a test should be developed for each adaptation of a test (Solano-Flores et al., 2009).

The multidimensionality of translation errors is an important principle of the TTE framework. Linguistic features of test items are interrelated and therefore result in multidimensional errors. For example, the misplacement of a comma may violate the writing conventions of a language and change the meaning of the sentence. Often this interrelationship creates tensions between how translators resolve errors, which may produce errors in other dimensions. Solano-Flores et al. (2006) applied the TTE framework to the Mexican adaptation of Third International Mathematics and Science Study (TIMSS-1995) and found that most of the items had errors in two or more dimensions, with an average of four dimensions. Based on this research, they recommend using a team of reviewers with a variety of expertise to address multidimensionality.

Another important feature of the TTE framework is the probabilistic space principle. This principle is based on the idea that a range of acceptable measurement error from test translation be defined by frequency and severity errors. The content and goals of the test can specify the range of acceptability with an intention to minimize rather than eliminate error.

Test translation theory of equivalence. From a linguistic academic discipline, another theory of test equivalence rooted in test theory and translation theory has emerged to address the

quality of international reading literacy assessment translations (Arffman, 2007). In translation theory, equivalence between the source and target versions of a test is evaluated through linguistic textual analysis. Different types of equivalence are prioritized in the translation of a test. For instance, connotative equivalence prioritizes style and register. Pragmatic equivalence seeks to attain equivalence of effect for the reader. Formal equivalence seeks to attain expressive or aesthetic equivalence. Textual equivalence focuses on similarity in information flow and cohesiveness.

Arffman (2010) used linguistic textual analysis to compare a PISA 2000 test in English and the translated Finnish version. Expository, narrative and non-continuous text types were examined according to syntax, lexis and text coding schemes. Of the six problem types that were identified, four were similar to those identified in previous research. One of the previously identified problem types was specialized terminology, which resulted in connotative differences. Arffman (2010) noted that there were some vocabulary differences between the Finnish and English versions that may have made the Finnish test easier. Finnish vocabulary may have been easier to understand because expressions were more popular and ordinary while the English version included more formal vocabulary. The two problem types that were not previously identified in the literature were specific to adaptation practices. The problem types were associated with strategies and choices translators made. Interference was identified as a common strategy used by translators. Translators use this strategy in an effort to follow the original text too closely or to improve upon the target text. Arffman (2010) found that when translators created texts that were too literal, this strategy resulted in lexical errors, imprecise language, ungrammaticalities and unintelligible language in the Finnish version (Arffman, 2010).

Arffman also found that the number and quality of linguistic bias differed between text types. The greatest numbers of problem types were found in the narrative and non-continuous text. The most frequent problem in the expository text was differences in word length, specialized and formal terminology, and interference by translators. In the narrative text, problem types tended to be associated with the possibility of many meanings for words, metaphors and personal pronouns in the English version. The English verb *cry* was provided as an example. In English, it can refer to shouting or weeping. In Finnish because there is not a similar word that shares dual meanings, a word with only one meaning was chosen. The most challenging problem in the non-continuous text was the compact language in the English text, which in some instances was impossible to convey in Finnish.

Consistent with the findings of Solano-Flores et al. (2009), Arffman concluded that it is impossible to attain full equivalence of difficulty between linguistic versions of international reading tests; however, it is possible to attain a high level of equivalence. Arffman (2012) also conducted a study in which she examined ways to improve upon expert reviews of language related differences. Reading literacy items translated from English to Finnish from PISA 2000 were reviewed through systematic linguistic comparisons of the two language versions by reviewers with linguistic and translation expertise. Of the two reading passages that were examined both were found to possess differences in grammar, syntax and meanings, which were attributed to inappropriate adaptation. Based on findings from this study, Arffman concluded that it is possible to improve expert reviews through the use of a more standardized procedure. Arffman makes a number of recommendations to develop more standardized procedures. Expert reviewers should provide a clear definition of the type of language-specific adaptation problem of an item. Further, testing organizations should provide examples of different types of

adaptation problems and their typical sources to aid in the development of clear definitions. Instructions on how to conduct the review, including how to rate the significance of language related differences should be provided. In addition, each country that participates in international LSAs should report results of national adaptation reviews to the respective national project centers (Arffman, 2012). Qualifications of reviewers should extend beyond expertise in the domain assessed to include experience with test development, knowledge of reading processes, response processes, and factors affecting item difficulty (Arffman, 2012). Finally, Arffman recommended the development of an expert review manual that outlines procedures, examples, a description of the method, instructions on how to rate the extent of adaptation problems and the necessary qualifications for conducting reviews.

Reading Literacy in a Sociocultural Context

In the field of language education, the question of how literacy tests are being used, the societal values that underlie those uses, their consequences and the ethical consideration of their use are being debated (Bachman, 2007; Bialystok, 2002; Cumming, 2009; Gee, 2001; Ochs & Schieffelin, 1995; Rueda, 2010). New literacy studies emphasize the importance of the sociolinguistic context of literacy proficiency beyond the structural level of a sentence (Bachman, 1987; Gee, 2001). These studies are based on philosophies of language from De Saussure, Derrida and Vygotsky. De Saussure (2011) argued that there is no inherent relation between words and objects rather language is a system of interdependent terms in which signifying actions correspond to conventions and social practices. More concisely, meanings of words depend upon their relations to other words. Likewise, Derrida's (1998) view of language was that words refer to other words and meaning relies on the context in which it is embedded and is dependent on social and historical practices. Meaning, according to Derrida, is a matter of

convention (Hymes, 1967). As human beings, we construct meanings and understandings as we engage in specific social interactions (Hawkins, 2004). With respect to the act of reading, Vygotsky theorized that it as an interactive process involving the use of cultural tools, symbols and texts (Hammerberg, 2004; Steiner et al., 1996). The theoretical frameworks of Saussure, Derrida and Vygotsky recognize that language development occurs within a culturally shared system through which individuals actively engage in customs specific to their communities.

Sociolinguistic theories describe the act of reading as a complex interplay between the content and context of the text and a reader's prior experiences, motivation, goals and sociocultural background (Alexander & Jetton, 2000; Hammerberg, 2004; Rumelhart, 1994; Ziegler & Goswami, 2005). Reading comprehension is an interactive process between reader, context and text, with the text providing the raw material for meaning making (Arffman, 2007). Readers participate in the creation of the text and its significance (Vera, 2011). When meaning is interactively constructed, comprehension involves negotiating a variety of possible meanings drawn from shared social and cultural experiences (Mesthrie, 2008). However, reading skills are integral to how meaning is generated. Reading skills are by-products of the reader's previous reading experiences, cognitive and metacognitive processes and social, cultural and educational contextual factors (Vera, 2011).

Given that the act of reading is a complex constructivist process and languages differ in form and meaning, establishing test equivalence across linguistic versions is complex but essential. Reading literacy research confirms that it is very difficult to adapt a test that is not culturally biased (Bonnett, 2002; Greenfield, 1997). Researchers attribute this difficulty, in part, to a fundamental assumption among test developers that competence in reading literacy is universal, independent of cultural and linguistic influences. But, many argue that when the

content and conventions of a culture are not represented in a test, the equivalence of difficulty is threatened (Arffman, 2010; Greenfield, 1997). When countries are ranked in comparative order of success, little attention is given to understanding the acquisition of competence with respect to cultural backgrounds. Yet, research that has grouped countries linguistically and culturally by the same items, has found clear distinctions and similarities (Bonnett, 2002; Ercikan et al., 2009; Grisay et al., 2007). For instance, France and Belgium pass and fail many of the same items, as do Britain and the U.S, while the profiles of France and Britain are quite different. Large construct differences have also been found with the Third International Mathematics and Science Study (TIMMS) scales between the U.S. and France (Ercikan et al., 2009).

Murata (2007) conducted a study that shows how sociocultural assumptions and prior knowledge effect perceptions and interpretations of text. The study examined how readers from different cultural backgrounds interpret news text. A reading comprehension questionnaire was administered to undergraduate Japanese students and non-Japanese undergraduate students in Britain and New Zealand. The study was based on the premise that language analysis or interpretation is based on readers' preconceived ideas and cultural assumptions. A topic that specifically creates radical differences in cultural attitudes between Japanese and non-Japanese students were chosen for the study. The chosen topic was whaling which is one of the most politically controversial subjects among animal rights and environmental protection movements. The text, taken from a British newspaper, had a strong anti-whaling tone and criticized Japan as a whaling nation that tried to abolish a whaling sanctuary. The researchers hypothesized that readers with anti-whaling cultural assumptions would be more likely to infer that the Japanese were responsible for a decrease in the whale population. Likewise, they hypothesized that readers with pro-whaling assumptions would have more difficulty processing information in the

text criticizing Japan as it conflicted with their prior knowledge and cultural assumptions. The results of the study showed that readers' from countries with anti-whaling cultural assumptions were more likely to blame Japanese society for the decline of the whale population. This study highlights the impartiality of interpretations and the ways in which readers are influenced by the interaction of their cultural assumptions, prior knowledge and their interpretations of a text. If readers were to take a test about whaling, cultural assumptions may affect their performances as a result of how text is interpreted and the inferences drawn.

Understanding language from a sociocultural perspective requires giving attention to language in use within communities, among individuals with cultural identities who are engaging in specific social functions (Hawkins, 2004). The meanings given to language are negotiated within situated cultural and language practices (Hawkins, 2004). Language creates and reflects specific social contexts and identities connected to social groups, cultures and historical practices. A sociocultural view of language recognizes that language does not exist as a unitary entity that is comprised of a collection of definable and consistent words and grammatical structures (Hawkins, 2004). From a sociocultural framework, we use language in ways that are specific to our social and cultural customs of engaging with language and texts. Sociocultural theories of literacy emphasize that meaning is constructed by what a reader brings to the text. This framework recognizes that literacy is neither content nor context free. Literacy is viewed as a complex cultural phenomenon in which social and cultural forces organize how readers use and understand text.

The sociocultural implications of language and meaning are inherently related to the acquisition of reading literacy and to test equivalence. For this reason, the *Standards* outline a framework to address the types of error that can occur when a test developed for one particular

group is adapted for a different linguistic and cultural group. Construct, method and item bias pose potential threats to test equivalence when a test developed for a particular cultural and linguistic group is adapted for a different cultural and linguistic group. It is essential to ensure that a test is measuring the same ability, trait or construct consistently for all groups of interest. The language that is used when a test is adapted for another group should reflect the social and cultural practices of that group. When the language used in an adapted test does not reflect the social and cultural customs of a group, the adapted version may contain awkward, unfamiliar or unnecessarily difficult words and sentences. These types of language problems may affect the meaning of words, the trajectory of examinees thinking processes and their responses to items. These problems create differences between test versions that may give rise to construct, method and item bias.

Summary

In this chapter, the uses and purposes of LSAs in a Canadian context were discussed. The increased reliance on LSAs to inform educational policy was considered with respect to the decentralized educational system in Canada, as well as concerns regarding consequential validity. The impact of bias on score comparability, including the influence of Western Anglo-Saxon assumptions on the development of instruments and implementation of psychometric practices was addressed. The establishment of criteria to ensure that appropriate methods are implemented when tests are adapted from source to target test versions were discussed with respect to the *Standards* and *Guidelines*. Quantitative and qualitative methods for examining construct and test equivalence such as CFA, IRT DIF and expert reviews were described, with research to substantiate their uses for this study. Two emerging theoretical frameworks to aid in the examination of test translation error were detailed to highlight the inherent linguistic and

cultural complexities of the adaptation process. To further underscore challenges to creating equivalent reading literacy tests and ensuring score comparability between linguistic groups, reader's perceptions and interpretations of texts was framed as a cultural phenomenon. Evidence suggests that although full equivalence across different language versions for international reading literacy tests is impossible, it is possible to attain a high level of equivalence (Arffman, 2010). The use of statistical methods allows for an examination of test equivalence at the item and scale level. Furthermore, recommendations to develop more standardized expert review procedures for examining test equivalence are available and evidence is being collected to support their use. In the next chapter, methods for conducting this study are described in detail.

Chapter III: Method

Overview

In this chapter, the methodologies that were used to examine test equivalence between the French and English language versions of PIRLS 2011 are described. The chapter begins with a brief account of classical test theory (CTT) analysis and item response theory (IRT) analysis. Aspects of PIRLS that are relevant to this study such as item characteristics, sample sizes, adaptation procedures, scoring and construct definitions are then described. Next, statistical and judgmental methodologies for examining item and test level comparability are discussed. The purpose of this study is to examine the following research questions:

1. To what extent is the factor structure of the construct evaluated by PIRLS 2011 equivalent across the French and English language groups?
2. To what extent do items function differently across the French and English language versions of PIRLS 2011 administered in Canada?
3. What do expert reviews reveal about the equivalence of French and English released items from PIRLS 2011 administered in Canada?

CTT and IRT Analysis

Item bias can be examined across linguistic versions of tests through classical test theory (CTT) analysis and item response theory (IRT) analysis. Both forms of analysis assume that each item measures some facet of a latent variable (e.g., reading ability) and that examinees possess a certain degree of that latent variable. Scores are considered to be indicators of unobservable latent variables. It is important to note that a latent variable is a statistical and mathematical variable created by the analyst to provide a predicted score based on examinees responses (Zumbo, 2007). The primary difference between CTT and IRT pertains to the unit of analysis.

With CTT, the unit of analysis is the sum of correct items. The test scale or the raw test score for each examinee is of primary interest. The disadvantage of CTT is that it does not account for how the psychometric properties of a scale vary as a function of the sample variance and the continuum of ability levels. With CTT, item characteristics and ability estimates are dependent on the sample. As a result, measurement error is the same for all examinees and is a function of the test length. Observed scores are comprised of a true score and an error score.

With IRT, the unit of analysis is individual items. Each examinee can be located along the continuum of an ability scale, based on the probability that an examinee answers an item correctly at each ability level. The probability is lesser or greater according to the examinee's ability level. Performance on a test item is a function of item parameters and ability level. An item response function (IRF) is a mathematical expression of the probability that an examinee at a given ability level with given item parameters answers a question correctly (Yen & Fitzpatrick, 2006). A graphical display of an IRF is an item characteristic curve (ICC). An ICC represents the probability of an examinee's response to an item on the y-axis and their underlying ability on the x-axis. Probability is plotted along an S-shaped curve, as a function of ability. The ICC is characterized by the parameters of the individual items.

In LSAs, such as PIRLS, a 3-Parameter Logistic IRT model is used to estimate the individual items. The first attribute is termed the 'b' parameter and represents the difficulty level of an item along the ability scale. In essence, it represents the amount of a latent variable that an examinee needs to answer an item correctly. More difficult items are located at the higher end of the ability level. The second property of an item is the discrimination index or the 'a' parameter, reflected by the incline of the slope. The discrimination index determines how well an item can differentiate between examinees at either side of the 'b' parameter. The third property of an item

is the ‘c’ parameter, also known as the guessing parameter. It is the likelihood that an examinee with a low ability level arbitrarily answers an item correctly.

With the capacity to estimate item parameters through ICCs, psychometric techniques were developed to assess differential item functioning (DIF). The statistical term DIF refers to techniques that compare ICCs of an item for different groups. If the probability of responding correctly to a single item differs for examinees of one group with equal proficiency levels compared to the other group, then the differences may reflect some characteristic of the item rather than true differences between groups. Typical characteristics that make an item biased against a language group may include word difficulty, cultural relevance, idiomatic expressions, equivalence of meaning, semantic differences, differences in the quantity and length of words, item format, item content and linguistic form. However, as previously mentioned psychometric differences between linguistic groups may also be due to other factors such as curriculum and cultural differences as well as the strategies and choices of translators. Differences in item responses are investigated after groups are statistically matched on the characteristic of interest. Test score differences between linguistic groups may reflect real discrepancies in knowledge and competencies; however, research has identified large amounts of DIF across different language versions of LSAs (Ercikan, 2002, 2003; Oliveri et al., 2011) that have been attributed to the incomparability of language versions (Ercikan et al., 2013; Oliveri et al., 2011).

Measure

The International Association for the Evaluation of Educational Achievement (IEA) is responsible for the development of the Progress in International Reading Literacy Study (PIRLS). The IEA was founded in 1959 to conduct comparative studies on educational practices and policies across countries. The IEA’s mandate is based on the premise that reading literacy is

the foundation for learning across all subjects, for personal and social growth and reading literacy equips children to participate fully in their communities. Key to this mandate is that the fourth year of schooling marks an important transition point in children's literacy development. Inaugurated in 2001 and administered in five-year cycles, the third PIRLS cycle was administered in nine Canadian provinces (IEA, 2012). PIRLS provides international comparative data on students' reading achievement after four years of primary schooling and measures trends over time. They recommend that if the average age of 4th grade students in a country is less than 9.5 years, administration of the test should be to the next highest grade. Reading achievement data from PIRLS has a scale average of 500 with a standard deviation of 100. In addition to assessing reading achievement, background information regarding students, home supports for literacy, teachers, schools and curriculum is collected.

In 2011, PIRLS was administered to students in their fourth year of formal schooling in 48 countries (IEA, 2012). The PIRLS 2011 database includes data from 334,446 students' worldwide. In Canada, approximately 23,000 students from 1,000 schools participated, with 16,500 writing the test in English and 6,500 in French. As mentioned previously, nine Canadian provinces participated: British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick French, Nova Scotia, and Newfoundland and Labrador. Alberta, Ontario and Quebec were benchmarking participants, which means that as a region they were categorized as distinct education systems with their own representative samples of students. Benchmarking participation allows regions to be regarded as separate countries so that students performances can be compared to that of all other participating countries.

Sampling approach. Students are sampled according to a two-stage stratified sampling design (IEA, 2012). With this design, schools that share common characteristics such as

geographic region or school type were sampled in the first stage and intact classrooms are sampled in the second stage. In regions where sizes were smaller, all schools and all grade 4 students were sampled. School level exclusions were permitted for schools that were geographically remote, those that had very few students, had a grade structure or curriculum drastically different from mainstream education or those that served primarily special needs students. Ministries of education in each province permitted student level exclusions for students that did not speak English or French and those with functional or intellectual disabilities.

Test design. In PIRLS 2011, items and passages are divided among booklets. PIRLS uses a matrix sample design in which ten 40-minute blocks are divided into 13 booklets. A block consists of a reading passage and 13 to 16 questions or items pertaining to the passage. The matrix design for PIRLS 2011 is shown in Table 2.1.

Table 3.1

Matrix Block Design by Booklet

Booklet	Block	Block
1	Flowers on the Roof	Fly, Eagle, Fly **(R)**
2	Fly, Eagle, Fly **(R)**	Shiny Straw
3	Shiny Straw	Empty Pot
4	Empty Pot	Leonardo da Vinci
5	Leonardo da Vinci	Day Hiking **(R)**
6	Day Hiking **(R)**	Sharks
7	Sharks	Where's the Honey
8	Flowers on Roof	Where's the Honey
9	Flowers on the Roof	Leonardo da Vinci
10	Fly, Eagle, Fly **(R)**	Day Hiking **(R)**
11	Shiny Straw	Sharks
12	Empty Pot	Where's the Honey
13	Enemy Pie **(R)**	Giant Tooth **(R)**

Note. **R** = Released items.

A matrix sample design is used when the purpose of an assessment is to measure the performance of groups rather than individuals. Subsamples of items in a booklet are assigned to subsamples of examinees. Data from booklets and examinees are aggregated to obtain a measure of group performance (AERA, APA & NCME, 1999). From the subset of items that examinees complete, performance on all the items is inferred. Thirteen booklets from PIRLS 2011 were administered in Canada (IEA, 2012). Each student completes one randomly assigned booklet that contains one literary passage and one information passage to form two 40-minute blocks. Six of the items and passages in PIRLS 2011 were from PIRLS 2001 and 2006. These six items and passages are referred to as trend tests because they are used to measure trends in reading literacy over time. Four new tests were developed for the 2011 assessment. Each test appears three times in one of the thirteen booklets to enable linking among booklets. The Reader booklet, which also contains one literary and one information passage, was administered as a separate booklet with 35 items and it is not linked with other booklets.

Scoring. PIRLS focuses on two reading purposes and four reading comprehension processes. The two reading purposes include reading for literary experience and reading to acquire and use information (CMEC, 2012). The four processes of comprehension targeted by PIRLS relate to how readers construct meaning from a text. The four processes of comprehension are a) focus on and retrieve explicitly stated information; b) making straightforward inferences; c) interpreting and integrating ideas and information; d) examine and evaluate content, language and textual elements (IEA, 2012). The processes of comprehension are discussed in the next section. There are a total of five literary and five informational passages or blocks, which are distributed across individual booklets (IEA, 2012). Booklets are designed to

provide an average of 15 score points that include 7 potential points from multiple-choice (MC), two or three short answer questions worth one or two points and one constructed-response (CR) question worth a maximum of three points. Approximately half of the questions are MC item-format with four response options and one correct option that is worth one point. The other half of the questions are constructed-response (CR) items worth one, two or three-points, depending on the depth of understanding that the question requires. Each CR question has a supplementary scoring guide that delineates the essential features of appropriate and complete responses. The scoring focus is on the student's ability to understand the text. To minimize the reading load of items, response options are written succinctly.

To ensure scoring consistency within each country designated IEA national research coordinators were required to randomly select 200 student responses to each CR item and have them scored independently by two scorers. The degree of agreement between the assigned scores is reported as a measure of reliability. To measure the reliability of scoring overtime between PIRLS cycles samples of administered and scored data from each country are submitted to IEA for rescoring before the scheduled scoring activity begins. If the agreement between scorers is less than 85%, scorers are required to retrain and previously entered scored are discarded and rescored. To ensure reliability of scoring across countries 100 of the randomly selected scored responses in English from each participating country were entered into a software program along with 100 scored responses from the previous PIRLS cycle (IEA, 2012).

Reading literacy. For PIRLS, reading literacy is defined as, **"**the ability to understand and use those written language forms required by society and/or valued by the individual" (Assessment Framework, 2009, p. 11).

According to the IEA, this definition reflects constructive and interactive processes described by theories of reading literacy. PIRLS examines two aspects of reading literacy, which include purposes for reading and processes of comprehension (IEA, 2012). Separate reading achievement scales were created in 2006 for reading for literary experience and reading to acquire and use information. The main form of text used by PIRLS for literary experience is fiction. The reader is perceived as one who engages with the text to become involved with imagined events, actions and consequences based on the readers' own experiences, knowledge and appreciation of literary forms. A wide range of text forms is used for each purpose of the reading section. To assess reader's ability to acquire and use information the text focuses on real aspects of the world. Informational text may not include headings or textual organizers and can be organized logically or chronologically. Information can be presented as continuous text or as brochures, lists, diagrams, graphs or advertisements and informational text and is often presented in more than one way.

For comprehension, each item refers to one of the following processes of reading: a) focus on and retrieve explicitly stated information; b) making straightforward inferences; c) interpreting and integrating ideas and information; d) examine and evaluate content, language and textual elements (IEA, 2012). The four processes of reading are assessed within each purpose for reading. Retrieving of explicit information requires not only retrieval but also how the information relates to the question. These questions are designed to require little interpretation or constructing of meaning. With the process of making straightforward inferences, the reader is typically required to connect two or more pieces of information but the connection between them must be inferred. The connection between the ideas is not explicitly stated but the meaning is intended to be relatively clear. With the comprehension process of

interpreting and integrating ideas and information, the reader draws on prior knowledge and information beyond what is provided in the text. Students need to integrate personal knowledge and experiences to with meaning in the text. The process of examining and evaluating content, language and textual elements requires that readers draw upon their knowledge of text genre, structure and language conventions. The extent of past reading experiences and familiarity with language usage are essential for this form of comprehension.

Separate scales were created for the four processes of comprehension. Text for comprehensive processes varies in length, syntactic complexity, abstractness of ideas, and organizational structure. For this reason, the nature of the text can have a considerable impact on the difficulty of the question.

Scaling methodology. An IRT psychometric model is used to analyze PIRLS assessment data (Foy, Brossman & Galia, 2012). The IRT model is a latent variable model that describes the probability of a student's response to an item on the basis of the student's underlying ability and the item's parameters. The application of IRT methodology entails estimating model parameters for each item. Reading literacy trends were measured over time by linking the 2011 assessment the 2006 assessment through the application of linear transformations. This process, referred to as concurrent calibration, places the results from each assessment cycle on the same scale. Concurrent calibration was achieved by retaining six passages and their items from 2006 for the PIRLS 2011 assessment cycle. Item parameters were estimated by combining 2011 assessment data with 2006 data (Foy et al., 2012). Once item parameters were estimated from 2006 and 2011 data, student's latent ability distributions for both assessments were estimated. The difference between these two sets of student distributions was the change in achievement from 2006 to 2011 assessment cycles. Next a linear transformation was found that transformed the data and

matched the distribution from 2006 to 2011. The linear transformation was then applied to the 2011 data scaled using the concurrent calibration. It is important to note that estimated item parameters from concurrent calibrations were based on all available item response data from each country. Estimated international item parameters were then used to create scores for each country. Student samples were weighted to ensure that each country contributed equally to the item calibration. This method of estimation assumes that international item parameters are representative of all countries.

With PIRLS data, there is a single scale for overall reading (Foy et al., 2012). However, proficiency scores are also generated for the subdomains, which include the purposes for reading and the processes of comprehension.

Adaptation. The French language version of PIRLS is developed through the forward translation of the English language version (IEA, 2012). With the forward translation model, a monolingual test developer constructs the test in a source language. Then translators adapt the test from the source language to the target language. Bilingual translators check the equivalence of the two tests. The advantage of forward translation as compared to other methods is it requires less time. The disadvantage is that the monolingual test developer is unable to judge the equivalence of the two tests. Guidelines were created by the IEA to assist in the translation of all the assessments materials. Each participating country is responsible for fitting PIRLS material to their cultural context. The guidelines recommend the preservation of the original information, the use of correct grammar and punctuation and preserving the meaning of idiomatic expressions rather than literal adaptations.

Procedure

To examine measurement equivalence assessed by PIRLS 2011 at the scale and item level both confirmatory factor analysis (CFA) and IRT methodologies were used. To evaluate the similarity of factor structures for reading literacy across French and English language groups CFA was conducted. The second methodology involves item level analysis using two DIF detection methods, which are discussed in the next sections. In the final stage, bilingual reviews of the four released passages were conducted to examine types of differences that may exist between the two language versions.

Methodology 1: Factor analytic invariance. Jöreskog (1971) introduced confirmatory factor analysis as a method to examine the relationship between responses to items and the latent variable the responses represent. It was designed to assess how well a hypothesized factor structure based on theoretical and empirical research fits the observed data prior to analysis. The score obtained on each item is considered to be a linear function of the latent variable and an error term (Zumbo, 2005). With CFA, factor loadings are forced onto the factor structure and then the fit of the model with the data is tested. The goal of CFA is to account for the covariation among the items by the latent variable. Scores that demonstrate consistent interrelations across linguistic groups provide evidence that the same latent variable is measured equivalently across test versions (Zumbo, 2003). Measurement invariance is essential to ensure the validity of scores across different groups (Kline, 2013).

Before CFA is conducted, the fit of reading literacy as one main factor underlying the data was tested. PIRLS was designed to have a single scale for overall reading (Foy et al., 2012). As previously mentioned, both a one and two-dimensional analysis were conducted to generate overall reading scores and subdomain scores. To ensure that the same latent construct is

measured with French and English language groups the data should fit one dominant factor model. To ensure that items function the same across groups factor loadings were examined for equivalence. Another reason for examining the factor structure with EFA is that IRT DIF methods are based on the assumption that the test is unidimensional. Exploratory factor analysis using the maximum likelihood estimation (MLE) method was conducted for each of the thirteen booklets with MPlus software (Muthén & Muthén, 1998). With the MPlus program, tetrachoric and polychoric correlation matrices can be used to estimate the relationship of dichotomous and ordinal data. Tetrachoric and polychoric correlation matrices enable dichotomous and ordinal data to be treated as if items have an underlying continuum of responses. The RMSEA statistic was used as the criterion for choosing the number of factors (Zumbo, 2003). The RMSEA fit statistic indicates how well the model with unknown but optimal chosen parameter estimates would fit the population's covariance matrix. Values of 0.05 are the standard criterion for the RMSEA fit statistic (Zumbo, 2003).

A simultaneous multi-group (by language) maximum likelihood CFA was conducted for each booklet of the thirteen booklets to test the fit of the hypothesized one factor structure for reading literacy. The purpose of a multi-group CFA is to examine the extent to which the factor loadings and error variances are invariant across the two or more groups (Breithaupt & Zumbo, 2002), in this case two language groups. The goal of CFA is to maximize the fit between the model and the data and minimize error. A full measurement invariance hypothesis was tested for equality of loadings and equality of uniqueness for language groups using tetrachoric and polychoric correlation matrices. The MLE method derives an estimation of the parameters based on whether the covariance matrix for the model is the same as the observed covariance matrix. The fit of these values are evaluated and then factor loadings are adjusted according to the lack

of fit of the model with the observed data (Russell, 2002). Fit indices indicate if the hypothesized factor structure fits the data.

Due to the influence of sample size on the chi-square goodness-of-fit test, alternative indicators of model fit that are less affected by variations in sample size are recommended. The root mean square error of approximations (RMSEA) and the root mean square residual (RMSR) were used to assess the fit of the model to the data. The RMSR is a measure of the average size of the residuals when the model is fit to the data (Zumbo, 2005). Values of 0.05 for both fit indices indicate close fit of a model to data. Sample size is also a consideration with respect to the normality assumption that underlies maximum likelihood estimation of CFA. The data were examined for violations of multivariate normality and local independence. A basic assumption underlying latent variable models is local independence. Factor analytic models are based on the assumption that the latent variable explains the observed covariances (Zumbo, 2003).

A CFA analysis only allows for the evaluation of construct equivalence at the test level. Zumbo (2003) found that construct variance at the item level is not evident at the test level. Zumbo recommends the use of CFA and DIF analyses to ensure the comparability of construct at the scale and item level (Zumbo & Koh, 2005). Researchers have found (Ercikan et al., 2005; Gierl et al., 1999; Sierci et al., 1998) that including CFA and DIF analyses resulted in comprehensive construct equivalence information.

Methodology 2: DIF analysis. The purpose of using DIF methods is to examine whether items function in the same manner for different groups. The detection of DIF across different linguistic versions of LSAs indicates that items are not functioning the same way across linguistic groups for examinees with equal abilities. In order to ensure that inferences regarding the performance of particular groups are valid, tests must yield comparable scores for the

comparison groups (Oliveri et al., 2012). When different groups of equivalent ability have the same expected probability of answering an item correctly, the item parameters are comparable and the test provides equivalent measurement.

With IRT analysis, the discrimination and difficulty parameters characterize the most important aspects of measurement equivalence with achievement tests. The discrimination parameter indicates how rapidly the ICC curve rises at the inflexion point. It represents the degree to which item response varies with ability level. The difficulty parameter indicates the location on the ability scale in which examinees have a 0.5 probability level of answering an item correctly. An item indicates DIF if response probabilities for examinees at the same ability levels depend on group membership.

In this study, the IRT-based Linn and Harnisch (L-H) parametric method was used to compute the difference in deciles between the predicted and observed probability of responding correctly to an item or of obtaining the maximum score. The predicted probability is based on a calibration using the combined group data and the observed probability is based on the minority group data. IRT parameters were calibrated using PARDUX software (CTB/McGraw-Hill, 1991). PARDUX software uses marginal maximum likelihood procedures to generate item parameters simultaneously for dichotomous and polytomous items. It can be used to estimate parameters with all the models used in this study. From the differences between the predicted and observed probabilities, a chi-square statistic is computed and converted to a Z-statistic. Items are flagged as DIF in favor of one language group or another according to a statistical significance level. Items with a Z-statistic ≥ 2.58 and $|p_{diff}| < 0.10$ are identified as moderate magnitude DIF, Level 2. Large magnitude DIF, Level 3 is identified by $|Z| > 2.58$ and $|p_{diff}| \geq 0.10$. A three-parameter logistic model was used to calibrate multiple-choice items and the two-

parameter partial credit model was used to calibrate constructed response items. A three-parameter response model provides an estimation of item difficulty ('b' parameter) and discrimination ('a' parameter) levels as well as a guessing ('c') parameter, which represents the probability of examinees with low ability levels answering an item correctly. A two-parameter response model does not include a guessing parameter. For this study, a two-parameter partial credit model (2PPC) is used, which is a special case of Bock's nominal model (Yen et al., 2006). This model is used to count the number of successfully completed steps when there is a range of potential scores in which higher scores on an item reflect greater ability (Yen et al., 2006). This model allows polytomous items to vary in their discrimination. The 2PPC is used for the short answer and CR items.

As methods for identifying DIF may not give identical results, the use of more than one method allows for the corroboration of DIF status for the items analyzed. The logistic regression (LR) and ordinal logistic regression (OLR) methods were the second method used (Swaminathan & Rogers, 1990). The LR and OLR method is based on the statistical modeling of the probability of responding correctly to an item, according to group membership, ability level and the interaction of these factors (Zumbo, 1999). The LR and OLR method can be used for binary and ordinal scoring formats. The LR and OLR method models the odds that an examinee will endorse a scale point for each point on the scale. The regression equation has more than one intercept coefficient but only one slope. The regression is conducted in a stepwise fashion with the item response as the dependent variable. The total score for each examinee is entered first followed by the grouping variable and then a group by total score interaction as the independent variables. The ordinal logistic regression (OLR) method is an extension of the LR method that was developed for polytomous items with ordinal data. With the OLR method, the model

predicts cumulative response probabilities of falling into or below thresholds across the number of response items minus one.

Statistical modeling provides a test of DIF based on the relationship between the item response and the total score by examining the effects of group membership for uniform DIF followed by the interaction effects for non-uniform DIF (Zumbo, 1999). Zumbo recommends examining both the Chi-square test and a measure of effect size to identify DIF with LR and to ensure that effects of DIF are significant. Effect sizes are a way of quantifying the size of the difference between the groups. The effect size statistic used in this study is R-squared, which indicates the proportion of shared variance between 2 or more variables. Items are classified as DIF in LR if the p value is less than or equal to 0.01. Items are identified as having negligible DIF if $R^2 < 0.035$, moderate DIF if $0.035 \leq R^2 \leq 0.070$, and large DIF if $R^2 > 0.07$ (Oliveri et al., 2012). By comparing the R^2 value of the grouping variable to the R^2 value of the total score the unique contribution attributable to language differences can be determined. The statistical significance of DIF is tested by subtracting the Chi-square value for the total score in step one from the Chi-square value of the interaction term in step three. The Chi-square value is then compared to the distribution function with 2 degrees of freedom testing for language and the interaction effects. One of the advantages of LR is that the test statistic provides an accompanying measure of effect size based on the difference of R-squared between the total score and the interaction term. When DIF statistical conclusions are based on both the p-value for the Chi-square difference test and the effect size criterion, the Type I error rate and statistical power is conservative (Zumbo, 2008).

Research has shown that with the LR method power is related to moderate and high item discrimination, unequal sample sizes and the interaction between these two factors. The power

for detecting non-uniform DIF with the LR method decreases significantly as item discrimination increases. Uniform DIF occurs when the difficulty parameters differ between the two groups. With uniform DIF, group differences on an item are the main effect and do not depend upon where an examinee scores on a latent continuum. Non-uniform DIF occurs when the ICCs cross and differences between the group responses on an item vary over the levels of the latent trait. Hence, the interaction of group by ability after statistically matching on the test score results is non-uniform DIF.

Evaluation of IRT model assumptions. The IRT model is based on three assumptions: a) essential unidimensionality; b) local independence; and c) model fit. If items measure one dominant continuous latent variable ranging from $-\infty$ to $+\infty$, the essentially unidimensional assumption is met (Hambleton et al., 1991). Performing EFA to evaluate the factor structure underlying the observed covariation among responses tests the unidimensionality assumption. The RMSEA statistic was used as the criterion for choosing the number of factors.

The assumption of local independence indicates that responses are not dependent on one another. It is assumed that relationships among the items are due to the conditional relationship with the latent variable (Hambleton et al., 1991). If there are associations among responses, parameter estimates may be inaccurate and items may appear as separate factors (Yen et al., 2006). In this study, the Q_3 statistic (Yen, 1984) is used to test this assumption. The Q_3 is a measure of the correlation between items in a test. Items are flagged as locally dependent when $Q_3 \geq .20$ (Yen, 1984).

The assumption of model fit is tested using a Q_1 fit statistic (Yen, 1984). The Q_1 statistic is a chi-square statistic that is used to test the null hypothesis of no statistical difference between the IRT model and the observed data. The Q_1 statistic is calculated by measuring the chi-square

distribution of all the standardized residuals for the different ability groups, with degrees of freedom equal to the number of ability groups minus the number of parameters in the model. The Q_1 statistic was standardized with $Z > 4.6$ indicating a poor fit (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998).

Methodology 3: Bilingual expert reviews. Both the *Standards* and *Guidelines* advise the inclusion of bilingual reviews as evidence to improve the accuracy of the adaptation process and to support verification of equivalence for different test versions. It is also important to identify DIF sources to understand how items affect the validity of interpretations for different groups and to determine how to minimize bias (Lazaraton & Taylor, 2007). The purpose of the review process is to identify differences between the original and translated test versions that may lead to differential response patterns between linguistic groups. Expert reviewers evaluate and rate the equivalence of items with respect to cultural and linguistic criteria.

In the last phase of this study, four bilingual reviewers with expertise in reading literacy and test construction experience evaluated released items from PIRLS 2011. They examined items in the two language versions to determine cultural relevance and equivalence of meaning, of overall format, of cues given to examinees to solve the problems and to ensure that the intended reading and difficulty levels are maintained (Bowles et al., 2008; Ercikan et al., 2013).

Selection and training of bilingual reviewers. For a review committee, it is recommended that a minimum of four individuals conduct blind item reviews (Ercikan et al., 2002; Ercikan et al., 2013). The expertise and experience of reviewers is fundamental to the review process. It is recommended that reviewers have the following expertise and experience: a) first language in the language of the source or target version of test; b) proficiency in both

languages; c) knowledge of the field of reading literacy; and d) familiarity with testing and test development (Arffman, 2012; Ercikan et al., 2013; Hambleton, 2005).

For this study, all the experts were fluent in both languages, with French as the first language for two of the reviewers and English as the first language for two other reviewers. Two of the experts have extensive experience with bilingual test development and test equivalence across French and English language versions and have expertise in reading literacy. The other two reviewers have elementary and middle school teaching experience. Expert reviewers compared the French and English language versions of four passages released from PIRLS 2011 and conducted linguistic and cultural reviews. A training session was conducted to ensure that reviewers understood the specific adaptation problems associated with the two types of reviews. In the training session instructions on how to conduct reviews and how to appraise and rate the significance of differences were discussed.

Training session. A training session was conducted based on previous research (Gierl & Khaliq, 2001; Ercikan, 2002) to describe the overall purpose of PIRLS and of this study and to explain the purpose and process of reviews. Reviewers were told that the primary purpose of the reviews was to examine items of the two language versions to identify any differences that may have led to performance differences for one language group. They were instructed to focus specifically on linguistic, cultural and format differences that may affect the equivalence of the tests. At the beginning of the session, reviewers were given copies of the four passages in French and English (see Appendix A), instructions and criteria for rating the significance of differences between test language versions (Appendix B), a checklist of potential translation errors (Appendix C), and worksheets to code errors. In order to familiarize reviewers with the types of linguistic, cultural and format differences examples from the checklist based on previous

research with expert reviews were discussed first (Ercikan & Lyons-Thomas, 2013). The group was then introduced to the rating criteria, which was adopted from a study that examined the comparability of French and English language versions of a LSA administered in Canada (Ercikan, 2002). Reviewers were asked to assign ratings between 0 and 3 for every item. They were instructed to give a rating of 0 if there were no linguistic, cultural and format differences between a French and English language item. Items that were identified as having differences between the two language versions were assigned a rating between 1 and 3, indicating the degree of the expected impact on student performance. Then reviewers described their understanding of the rating criteria. Finally several examples of linguistic and cultural differences were selected to demonstrate how to appraise and rate the significance of differences between the French and English language versions.

In the first stage, items were reviewed independently without knowledge about which items were identified statistically as DIF. In the second stage, the group only reviewed items that were identified statistically as DIF for which there were rating differences among reviewers. For the first stage, reviewers were instructed to independently examine and compare passages and items from the two language versions simultaneously. If differences were found, reviewers were instructed to rate the degree of differences between the two language versions according to the rating criteria in Table 3.1 and to determine if those differences would lead to performance differences between the two language groups. They were then instructed to describe the source of the difference on the coding sheet and to then refer to the checklist of potential problems to identify the specific type of linguistic and cultural difference. If they identified several different types of differences, they were asked to fully describe these differences on the coding sheet. After completing that step, they were asked to consider if the difference favored one linguistic

group over another and to record the group it favored on the coding sheet. If they had recorded several differences with an item of which some had favored one group and some favored another, they were asked to include an explanation. Finally reviewers were instructed to indicate the degree to which they were confident of their rating for each item on a scale of 0 to 3, with 0 indicating they were not confident and 3 indicating they were very confident. They were instructed to record any additional comments or differences they identified on the coding sheet.

Table 3.2

Expert Review Rating Criteria

Rating		Meaning Associated with Rating
0	No difference	No difference between the two versions
1	Negligible Difference	Minimal difference between the two versions
2	Somewhat different	There are clear differences between the two versions but they are not expected to lead to differences in performance between the two groups of examinees.
3	Very different	There are clear differences between the two versions that are expected to lead to differences in performances between the two groups of examinees

In the second stage of the review process, only items that were identified statistically as DIF that were rated differently by reviewers were discussed as a group. The group discussion consisted of each reviewer saying how they had rated an item, describing the differences they identified and explaining they're rational for a rating. After each reviewer had an opportunity to

explain their rating of an item, there was a group discussion about the nature and degree of differences with respect to the rating criteria. The discussion continued until they reached a consensus about a rating.

Summary

The chapter opened with a brief overview of the differences between CTT and IRT analysis. The framework and technical details of PIRLS were described, as well as the adaptation procedures and the conceptualization of reading literacy. Statistical procedures to be conducted for this study including CFA and DIF detection methods at the item and test level were summarized. The review process was described and the training session was outlined. In the next chapter, results of the statistical and expert review analysis are detailed in reference to the degree of comparability between the French and English language versions of PIRLS 2011 and the potential sources of incomparability.

Chapter IV: Results

In this chapter results of the study are presented. The purpose of the study was to answer the following research questions.

1. To what extent is the factor structure of the construct evaluated by PIRLS 2011 equivalent across the French and English language groups?
2. To what extent do items function differently across the French and English language versions of PIRLS 2011 administered in Canada?
3. What do expert reviews reveal about the equivalence of French and English released items from PIRLS 2011 administered in Canada?

Descriptive statistics of the sample and the organization of the booklets are presented first, followed by descriptive statistics of the data. Next, results of simultaneous multi-group CFA for each booklet including values for RMSEA and RMSR fit indices are described to address the question of the extent to which factor structures for reading literacy are equivalent across language groups. To address the extent to which items function differently across language groups results from the LH IRT method are detailed including an evaluation of IRT model assumptions. Then results for the LR and OLR DIF methods are compared to the IRT DIF method. In the final section of this chapter, reviews by four expert reviewers of four passages and accompanying items are detailed and compared to items identified as DIF by statistical methods.

Description of Sample

Thirteen booklets were administered to approximately 23,000 students, of which about 16,500 wrote the test in English and 6,500 wrote the test in French. Participating Canadian provinces included British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New

Brunswick French, Nova Scotia, and Newfoundland and Labrador. Alberta, Ontario and Quebec were benchmarking participants, allowing those provinces to compare their test scores to those of other countries. In this study, data from all participating provinces were included in the analysis. In order to ensure reliable comparative analysis between the two language groups, a random sample of 500 English language students per booklet were selected using SPSS software to approximate the average sample size of approximately 430 for the French language group for twelve of the booklets. The French groups include all students who received the booklets in French. It is important to note that drawing a random sample may not produce a representative sample for the four provinces that were oversampled. With oversampling, there is a greater likelihood of drawing subjects from the oversampled provinces. The use of sampling weights could have addressed the problem of oversampling but there are significant limitations to using sampling weights with matrix designs. The sample size for both language groups was larger for Booklet 13, referred to as the 'Reader Booklet'. Having similar sample sizes across the two groups is recommended with factor analysis, and DIF detection methods because these methods are sensitive to sample size, which may affect the comparisons between groups. Larger sample size increases the power of statistics. Table 4.1 shows sample size by language and booklet. On the whole, samples were evenly distributed by gender with a slightly higher representation of males in many of the booklets.

Table 4.1

Sample Sizes by Booklet

Booklet	French Total	English Total
1	437	500
2	433	500
3	425	500
4	432	500

Booklet	French Total	English Total
5	427	500
6	429	500
7	423	500
8	426	500
9	425	500
10	433	500
11	428	500
12	433	500
13	681	726

Reading Assessment Data

Twelve of the 13 PIRLS booklets administered in Canada contained ten overlapping blocks or passages. Each booklet contains two blocks. For instance, the two blocks in Booklet 1 are 'Flowers on the Roof' and 'Fly, Eagle, Fly.' The two blocks in Booklet 2 are 'Fly, Eagle, Fly' and 'Shiny Straw.' The organization of the 13 booklets is shown in Table 4.2. The thirteenth booklet is referred to as a 'Reader Booklet' and contains distinct passages and items that do not overlap with the other booklets. Each booklet contains the two reading domains, literary and informational passages, that PIRLS assesses and the four processes of comprehension.

Table 4.2

Organization of Booklets

Booklets	Questions in Booklet	Number of Questions	Maximum Score
1	Flowers on the Roof Fly, Eagle, Fly **(R)**	25	32
2	Fly, Eagle, Fly **(R)** Shiny Straw	26	35
3	Shiny Straw Empty Pot	34	42
4	Empty Pot Leonardo da Vinci	32	40

Booklets	Questions in Booklet	Number of Questions	Maximum Score
5	Leonardo de Vinci Day Hiking **(R)**	24	32
6	Day Hiking **(R)** Sharks	24	32
7	Sharks Where's the Honey	28	37
8	Flowers on Roof Where's the Honey	29	36
9	Flowers on Roof Leonardo da Vinci	25	33
10	Fly, Eagle, Fly **(R)** Day Hiking **(R)**	24	31
11	Shiny Straw Sharks	26	36
12	Empty Pot Where's the Honey	36	43
13	Enemy Pie **(R)** Giant Tooth **(R)**	35	42

Note. **R** = Released items.

There were differences in mean raw scores for the two language groups for all thirteen booklets. Mean and standard deviation differences are shown in Table 4.3. Independent samples t-tests were conducted to compare total mean scores for French and English language groups. There were significant differences in mean scores for all thirteen booklets at the $p < 0.01$ level. For twelve booklets, mean score differences ranged from 1.80 to 3.82. The greatest mean difference between language groups was 6.07 for Booklet 13. Mean scores for the English language groups were significantly higher across all thirteen booklets, with French language students averaging 3.08 points lower across thirteen booklets. The maximum possible scores per

booklet can be found in Table 4.2. Standard deviations for all of the booklets for both languages ranged from 5.16 to 9.19, which indicates that there is a large total score variation within each language group. Percent correct values by booklet for each language group are displayed in Appendix D.

Table 4.3

Descriptive Statistics for Booklets 1-13

		French		English		
Booklet	# of items	N	Mean (SD)	N	Mean (SD)	Difference
1	25	437	19.56 (5.83)	500	22.87 (5.60)	3.31*
2	26	433	21.18 (7.13)	500	23.24 (7.30)	2.06*
3	34	425	25.11 (9.02)	500	27.76 (9.03)	2.65*
4	32	432	22.42 (7.78)	500	25.62 (7.30)	3.21*
5	24	427	16.36 (5.80)	500	19.45 (5.83)	3.09*
6	24	429	18.26 (6.41)	500	20.06 (6.44)	1.80*
7	28	423	18.78 (7.77)	500	21.80 (7.98)	3.02*
8	29	426	19.54 (7.32)	500	23.02 (7.46)	3.48*
9	25	425	17.74 (5.66)	500	20.87 (6.12)	3.13*
10	24	433	19.32 (5.77)	500	21.32 (5.56)	2.00*
11	26	428	20.77	500	23.12	2.34*

		French		English		
Booklet	# of items	N	Mean (SD)	N	Mean (SD)	Difference
			(7.50)		(7.75)	
12	36	433	23.02	500	26.84	3.82*
			(9.30)		(9.19)	
13	35	681	19.64	731	25.71	6.07*
			(8.95)		(9.51)	

Research Question One

Factor structure analysis.

The premise of validity in cross-cultural and cross-language group comparisons requires that test scores measure the same construct on the same metric. Score differences across language groups may be due to a variety of factors; therefore, there must be evidence to demonstrate that constructs are comparable. Measurement invariance can be examined by conducting CFA to test a hierarchy of hypotheses with increasing levels of equality constraints across groups. The least constrained model to test is configural invariance. In the configural model the number of factors and the pattern of the loadings are constrained to be the same across groups. Configural invariance allows for a basic investigation of how similar the data structures are across groups. If configural invariance is met, weak or metric invariance is the next level of measurement invariance to test. At the weak invariance level, the equality of item-factor scores are statistically tested to examine if factor loadings are equivalent across groups. Weak invariance indicates that the strength of the relationships between a factor and items are equal across groups and the unit of measurement of the latent variable is the same across groups for all the items. The next level of measurement invariance is strong or scalar invariance. Strong measurement invariance is obtained if the intercepts are equal for both groups across all the items, which means that items have the same point of origin across groups. For instance, a score

of two in the English language group may be equal to a score of one in the French language group. The regression line for predicting a score based on ability level should be the same for both language groups. When this level of invariance is achieved, it means that factor means can be compared across groups. Strict invariance is the highest level of invariance. At this level, residual variances for all the items are equal across groups. The residual variance is the portion of item variance that is not attributable to the factor. Residuals are assumed to be conditionally independent random errors. If the conditional independence assumption holds, item residuals do not correlate with those of other items or with common factors after conditioning on the factor score.

Failure to demonstrate configural invariance indicates that the construct of interest differs across groups. If the null hypothesis is rejected based on the χ^2 statistic for configural invariance, this indicates that the pattern of fixed and free loadings are not equivalent across groups. This means that there are differences between groups in how well the construct captures the items. The decision rule for whether MI holds at each level depends on a combination of indicators. The most common statistic used to evaluate MI is the Chi-square difference between two models; however, as previously mentioned, large sample sizes can result in high levels of Type I error, which means that the null hypothesis is rejected when it shouldn't be. A more conservative approach to examine results for multi-group CFA was used in this study because the use of multiple criteria are recommended to address the limitation of the χ^2 goodness of fit statistic (Zumbo, 2007). Fit indices are suggested because they are relatively unaffected by variations in sample size when testing models. A two criteria strategy is recommended to evaluate model fit. The RMSR fit statistic has been examined in several Monte Carlo studies and has been found to be sensitive to the mis-specification of factor models (Russell, 2002). It is considered a more

stringent criterion. The RMSEA fit statistic is recommended for the evaluation of configural invariance (Zumbo, 2007). The recommended criteria for a close fit for both the RMSEA and RMSR vary between 0.05 and 0.06 (Wu et al., 2007; Zumbo, 2007).

In this study, a simultaneous multi-group CFA for a single factor was conducted. The purpose of a multi-group CFA is to examine the extent to which the model parameters are invariant across the French and English language groups. Configural invariance was the first level of measurement invariance tested to determine if the overall pattern of the model is equivalent across linguistic groups. Chi-square, RMSEA and RMSR results from a multi-group CFA single factor analyses for Booklets 1-13 for the two language groups are shown in Table 4.4. The χ^2 goodness of fit statistic was significant for all thirteen booklets. Three of the thirteen booklets that did not meet the χ^2 criteria did meet the RMSR and RMSEA criteria. Ten of the booklets did not meet two of the three criteria for a good fit and six failed to meet all three criteria. For this reason, the next level of measurement invariance was not tested.

Table 4.4

Fit Indices for Configural Invariance of Booklets 1-13

Booklet	N	χ^2	p	RMSEA	RMSR	CI
1	929	1493.155	**0.00**	0.057	**0.072**	
2	924	1325.501	**0.00**	0.048	0.059	X
3	919	11346.245	**0.00**	**0.141**	**0.066**	
4	921	8370.668	**0.00**	**0.127**	**0.067**	
5	919	1098.435	**0.00**	0.047	0.060	X
6	909	1165.541	**0.00**	0.050	**0.066**	
7	912	6652.516	**0.00**	**0.131**	**0.079**	
8	917	7396.484	**0.00**	**0.133**	**0.083**	
9	918	1383.234	**0.00**	0.053	**0.068**	
10	926	1172.979	**0.00**	0.049	0.061	X

Booklet	N	χ^2	p	RMSEA	RMSR	CI
11	919	1169.983	**0.00**	0.042	**0.072**	
12	925	20605.498	**0.00**	**0.182**	**0.072**	
13	1398	24209.729	**0.00**	**0.167**	**0.083**	

Note. χ^2 , RMSEA and RMSR rejection of configural invariance are highlighted in bold. CI = configural invariance. Booklets that met configural invariance are marked by X.

These results indicate that there are differences between the two language versions for a one-factor model. Although one primary factor accounted for a large portion of the variance for both languages in all thirteen booklets, EFA results suggest that there were differences in the structural patterns and equivalence of the data. Results from EFA analysis to assess essential unidimensionality are discussed in the next section of this chapter. However, for the purpose of the present discussion on factor differences across the two language versions, EFA results from some of the booklets are included in this section. As previously mentioned, results of the EFA analysis indicate that there were differences between the two language versions regarding the number of factors that best represent the data. For instance, examination of the factor structure from one of the three booklets that met the RMSEA and RMSR criteria for configural invariance indicates that there are clear differences. Results for the EFA analysis from Booklet 2 are displayed in Table 4.5. The RMSEA value of 0.044 indicates that a one factor model is a better fit for the English language version than it is for the French language version, RMSEA = 0.050.

Table 4.5

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 2

Factors	English				French			
	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA
1	5.948	70.04	70.04	.044	6.373	70.16	70.16	.050
2	1.308	15.40	85.44	.040	1.467	16.15	86.31	.041
3	1.236	14.55	100.00	.036	1.244	13.69	100.00	.037

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

Furthermore, items in Booklet 2 did not result in significant nonzero loadings on one factor for both language groups. In an examination of factor loadings, values > 0.30 are often considered significant (Brown, 2006). Loadings for the twenty-six items in Booklet 2 differed across language groups. Item 11 did not have a significant nonzero loading for the English language group but it did for the French language group. Item 19 did not have a significant nonzero loading for the French language group but it did for the English language group. These differences are supported by the RMSEA values of the EFA in Booklet 2. For the English language group, the RMSEA value was 0.044, which indicated that a one factor model was a better fit than it was for the French language group, RMSEA = 0.050. A two factor model appears to represent the data better for the French language group, RMSEA = 0.041. For comparisons of loadings for items in Booklet 2 refer to Table 4.6.

Table 4.6

Estimated EFA Factor Loadings Booklet 2

Item	French	English	Difference
	Estimated factor loadings	Estimated factor loadings	
1	0.314	0.419	0.105
2	0.398	0.405	0.007
3	**0.258**	**0.174**	0.084

Item	French	English	Difference
	Estimated factor loadings	Estimated factor loadings	
4	0.458	0.415	0.043
5	0.527	0.384	0.143
6	0.410	0.499	0.089
7	0.404	0.312	0.092
8	0.501	0.415	0.086
9	0.339	0.365	0.026
10	0.395	0.499	0.104
11	0.412	**0.184**	0.228
12	0.520	0.487	0.033
13	0.392	0.377	0.015
14	0.556	0.452	0.104
15	0.415	0.363	0.052
16	0.484	0.479	0.005
17	0.530	0.426	0.104
18	0.543	0.519	0.024
19	**0.224**	0.459	0.235
20	0.553	0.520	0.033
21	0.634	0.686	0.052
22	0.417	0.335	0.082
23	0.517	0.399	0.118
24	0.543	0.641	0.098

Item	French	English	Difference
	Estimated factor loadings	Estimated factor loadings	
25	0.577	0.603	0.026
26	0.492	0.418	0.074

Note. Items without significant nonzero factor loadings are highlighted in bold.

There were factor structure discrepancies for many of the booklets. For instance, as shown in Table 4.7 a four- factor structure fails to represent the data for the French language version but does represent the data well for the English language version for Booklet 5.

Table 4.7

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 5

Factors	English				French			
	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA
1	5.154	59.08	59.08	.043	5.020	66.59	66.59	.044
2	1.271	14.57	73.65	.038	1.345	17.84	84.43	.038
3	1.153	13.22	86.87	.035	1.174	15.57	100.00	.034
4	1.146	13.14	100.00	.030				

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

The RMSEA values displayed in Table 4.8 for Booklet 8 show that a two-factor model is a better fit for the English language version, while a two or three factor model fits the French language version. Such discrepancies suggest that there are differences in factor structures between the two language versions.

Table 4.8

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 8

Factors	English				French			
	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA

1	6.831	80.20	80.20	.084	6.099	67.32	67.32	.082
2	1.686	19.80	100.00	.043	1.612	17.79	85.13	.044
3					1.348	14.88	100.00	.038

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

As shown in Table 4.9, the first factor accounted for 69% of the variance for the English language version but only 62% for the French language version in Booklet 9.

Table 4.9

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 9

	English				French			
Factors	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA
1	5.590	68.53	68.53	.045	4.503	61.99	61.99	.050
2	1.322	16.21	84.74	.039	1.447	19.92	81.90	.043
3	1.245	15.26	100.00	.034	1.315	18.10	100.00	.038

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

In Booklet 10, as shown in Table 4.10 a four-factor structure fails to represent the data for the French language version but does represent the data for the English language version.

Table 4.10

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 10

	English				French			
Factors	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA
1	4.899	56.22	56.22	.045	4.768	64.21	64.21	.049
2	1.337	15.34	71.56	.039	1.372	18.48	82.69	.043
3	1.298	14.90	86.46	.032	1.286	17.32	100.00	.038
4	1.180	13.54	100.00	.027				

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

As shown in Table 4.11, one primary factor represents 83% of the variance for the English language version and 69% for the French language version and a three-factor model only represents the data well for the French language version.

Table 4.11

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 11

	English				French			
Factors	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA
1	6.533	83.17	83.17	.045	5.737	69.14	69.14	.046
2	1.322	16.83	100.00	.038	1.278	15.46	84.87	.042
3					1.250	15.12	100.00	.038

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

As shown in Table 4.12, a one-factor model accounts for 92.37% of the variance for the data in the English language version and 80.02% for the data in the French language version and the RMSEA values are very different for all the factor structures between the two language groups.

Table 4.12

Eigenvalue, Variance and RMSEA Results for Factors in Booklet 13

	English				French			
Factors	Eigen-value	% of variance	Cumulative %	RMSEA	Eigen-value	% of variance	Cumulative %	RMSEA
1	9.237	92.37	71.88	.067	8.002	80.02	69.04	.083

Note. RMSEA = Root mean square of approximation. RMSEA ≤ .05 indicates good fit.

Summary

Overall results from multi-group CFA and from EFA provide evidence for factor structure differences throughout all thirteen booklets. Of the two booklets that met two of the fit criteria from multi-group CFA, results from EFA indicate that there are differences in these booklets across language groups. For instance, EFA results from Booklet 5 indicate that a four-

factor model represents the data well for the English language group but does not adequately represent the data for the French language group. In an examination of EFA results from the third booklet that met two of the criteria for evaluating multi-group CFA equivalence, there were also differences that indicated different factor structures. For instance, the EFA estimated factor-loading output for Booklet 10 show that four items from the French language version fail to result in significant nonzero loadings compared to three items from the English language version. The factor loading for item 24 in Booklet 10 is 0.504 for the French language group and 0.282 for the English language group. The factor loading for item 3 in Booklet 10 for the French group is 0.192 and 0.337 for the English language group. Overall results for the simultaneous multi-group CFA indicate that there are significant differences between how well a one factor structure fits the data for ten of the thirteen booklets.

Research Question Two

Differential item functioning.

Item response theory analysis. An IRT model expresses the association between responses to items and the latent ability measured by a test. The IRT-based L-H model assesses the fit of the model for the target group using item ability parameter estimates. The probabilities for both language groups are based on item parameter estimates for the combined sample. The probabilities from the two language groups are then compared using observed proportion correct statistics.

Evaluation of IRT model assumptions. In order to ensure that an IRT model is a good fit for the PIRLS 2011 test it is essential to examine evidence of fit between the data and an IRT model. A good fit between the data and model can be examined by checking the principal assumptions underlying unidimensional IRT models. Several assumptions were examined in this

study. To assess whether the PIRLS 2011 booklets are sufficiently unidimensional to apply an IRT model EFA results for all thirteen booklets for each language group are examined. To follow is an assessment of item fit using the Q_1 statistic for all the items in the thirteen booklets. The Q_1 statistic entails dividing the IRT ability scale into 10 score cells for the two language groups and then the observed and expected proportions of student's getting the item right are calculated and compared. The final measure of IRT model-data fit that was examined in this study is local item independence. The local independence assumption is based on the notion that observed items are independent of one another given a score on the latent variable. This postulation is important because it assumes that the latent variable accounts for the associations between the observed items.

Unidimensionality. A unidimensional IRT model assumes that differences among students or items are due to a single ability or achievement captured by the test. The student's level of achievement is measured by the test and reflected in scores. A condition of this assumption is that all the items measure the same latent ability. The relationship between ability level and true score is based on a nonlinear increasing relationship. There is an expectation that the parameters that characterize an item are not dependent on the ability distribution of students and the ability parameters of students are not dependent on the test items (Hambleton et al., 1991). To meet this assumption requires evidence of essential unidimensionality, which necessitates that the first factor should account for a large share of the variance.

Exploratory factor analysis was conducted for each language group by booklet using maximum likelihood estimation (MLE) with MPlus software. Factor analysis models were estimated using binary and ordinal variables in the model. As previously discussed, EFA creates a statistical model of relations between a set of variables. The intention of EFA is to explore

factor alternatives, as there are numerous models to fit the data. With strict unidimensionality a single factor model should account for the associations between items and there should be no secondary minor dimensions (Slocum-Gori & Zumbo, 2011). With essential unidimensionality a test is dominated by a single latent factor but includes secondary minor latent factors (Slocum-Gori & Zumbo, 2011). An indication of essential unidimensionality is if the first factor accounts for a substantial part of the matrix variance (Lord, 1980). There are a number of methods to test for essential unidimensionality. Zumbo recommends a number of criteria to examine essential unidimensionality including an investigation of eigenvalues. The MLE RMSEA statistic was found to be a promising approach but requires more empirical evidence. In this study, the eigenvalue criteria were used to determine essential unidimensionality, with the first factor accounting for a substantial part of the matrix variance. The ratio of first to second eigenvalues greater than three rule was used. The RMSEA statistic was also included to indicate the fit of various factor models.

Exploratory factor results indicate that one main factor represents a large share of the variance for both the English and French language versions across all 13 booklets, as shown in Table 4.13. These results suggest that the test is sufficiently unidimensional to apply an IRT model to the data. For Booklet 1, a primary factor accounts for 50% of the variance in the English language version and 48% for the French language version. The RMSEA statistic indicates how well the model fits the population covariance matrix with unknown and optimal parameter estimates. The RMSEA statistics also indicate that a single dominant factor represents the data well for both language groups. A primary factor also accounts for a substantial proportion of the variance for both language groups for booklet 2, representing 78.60% of the variance for the English language group and 81.24 % of the variance for the French language

group. In Booklet 4, one factor accounts for 66 % of the variance in the English language version and 72% in the French language version. As shown in Table 4.13, a one-factor model accounts for a substantial proportion of variance for both languages across all the booklets. Variance percentages for a one-factor model were particularly high for Booklets 12 and 13, accounting for a minimum of 80%.

Table 4.13

Summary of Percent Variance Accounted by First Factor by Booklet

	English			French		
Booklet	Eigenvalue	% of variance	RMSEA	Eigenvalue	% of variance	RMSEA
1	4.997	50.00	.048	4.798	48.00	.049
2	5.948	59.50	.044	6.373	63.73	.050
3	7.860	78.60	.054	8.124	81.24	.063
4	6.567	65.67	.061	7.224	72.24	.062
5	5.154	51.54	.043	5.020	50.20	.044
6	5.369	53.69	.049	5.247	52.47	.044
7	6.686	66.86	.081	6.233	62.33	.072
8	6.831	68.31	.084	6.099	60.99	.082
9	5.590	55.90	.045	4.503	45.03	.050
10	4.899	49.00	.045	4.768	47.68	.049
11	6.533	65.33	.045	5.737	57.37	.046
12	8.472	84.72	.061	7.956	79.56	0.57
13	9.237	92.37	.067	8.002	80.02	.083

Item fit. The Q_1 statistic was used to determine fit of the test items to the IRT models. Items with fit statistics $Z > 4.60$ indicate a poor fit at alpha = 0.05 level. Items may be identified as a poor fit if they do not accurately predict the performance of a subgroup (Yen & Fitzpatrick, 2006). A goodness of fit statistic for IRT is sensitive to sample size because statistical power is associated to significance tests that may be too low to detect model-data discrepancies (Hambleton, Swaminathan & Rogers, 1991). Parameter estimates from smaller sample sizes are also more likely to have larger standard errors. A number of factors other than sample size such as problems with item quality can also cause apparent misfit; therefore, the recommendation is to regard statistical analysis of item fit as informative rather than conclusive (Yen & Fitzpatrick, 2006). There were no poor fit items within the 13 booklets, indicating that the items fit the IRT models well.

Local item dependence (LID). The LID assumption is fundamental to latent variable models such as factor analysis and latent trait models such as IRT analysis. With this assumption, there must be evidence that the relationships or covariations between observed items explain the latent variable. Violation of the local independence assumption suggests that the association between observed items is not fully explained by the latent variable and other factors influence responses of students on some items. Potential causes of LID include but are not limited to factors such as student fatigue, speededness, practice, special item formats, a shared stimulus or passage, item chaining, item redundancy, multidimensionality and differential opportunity to learn (Yen & Fitzpatrick, 2006). In previous research LID has been found among items related to a common passage with a minimal effect. It has also been found with CR items in which students must explain their reasoning underlying their answers to previous items. The greatest effect of LID is that standard errors of test scores may be under-predicted (Yen &

Fitzpatrick, 2006). Positive LID indicates that on the basis of student's total scores and ability levels students perform either better or worse than expected on one item and another item. Negative LID indicates that students perform better than expected on one item and worse than expected on another. As previously mentioned, LID may occur when the IRT model does not fully explain relationships between item scores, which may suggest that the dimensionality of a test is defined by other abilities. The Q_3 statistic (Yen, 1984) was used to evaluate correlations between performances on two items given the model specifications based on ability estimates. Values greater than 0.20 indicate LID.

The Q_3 statistic results from Booklets 1, 2, 5, 6,10 and 11 indicate that all the items meet the assumption of local independence.

Pairs of items flagged as LID for Booklets 3 and 4 are listed in Table 4.14. The Q_3 values range from 0.20 to 0.50. The number of LID pairs in 'The Empty Pot' passage suggests that something other than overall reading ability level accounts for student's performances on these items in this passage. The same LID pairs from 'The Empty Pot' passage were identified in Booklets 3 and 4. An examination of EFA results for items 31-34 in Booklets 3 and 4 indicate that loadings on a one-factor model are similar for these items, which may mean that they require similar types of reasoning processes. For instance, item 32 has a factor loading of (0.519), item 33 has a factor loading of (0.522) and item 34 has a factor loading of (0.531). In addition, these items are all designed to test the reading comprehension process of interpreting and integrating ideas and information. Items 31-35 are also labeled in the PIRLS coding manual as a chain of connected items, which means the items are not independent from one another. Items that are connected are another potential cause of LID.

Table 4.14

Item Pairs with Local Item Dependence Booklets 3 and 4

Booklet	Passage	Item Pair	Q_3 Value
3	The Empty Pot	31 and 32	0.496
3	The Empty Pot	31 and 33	0.360
3	The Empty Pot	31 and 34	0.336
3	The Empty Pot	32 and 33	0.222
3	The Empty Pot	32 and 34	0.233
4	The Empty Pot	17 and 18	0.449
4	The Empty Pot	17 and 19	0.304
4	The Empty Pot	17 and 20	0.441
4	The Empty Pot	18 and 19	0.287
4	The Empty Pot	18 and 20	0.251

Pairs of items with Q_3 values > 0.20 from Booklets 7 and 8 are listed in Table 4.15. The Q_3 values range from 0.20 to 0.516. The negative LID pairs in Booklet 7 and 8 indicate that students' perform better than expected on one item and worse than expected on another according to their ability level, suggesting that another factor may account for their responses to items. Items 14 and 19 are both designed to assess the reading comprehension processes of interpreting and integrating ideas and information. Students that did well interpreting and integrating ideas and information for one of these items did poorly on the other item. The LID pairs 19/20 and 19/21 in Booklet 7 and 20/21 and 20/23 in Booklet 8 are numbered differently but are the same item pairs. Items 19-22 from the passage 'Where's the Honey?' are all designed to assess the same reading comprehension processes; therefore, they were either expected to perform better or worse on these pairs of items. Another potential cause of LID for items 19-22 is that they form a cluster of related items.

Table 4.15

Item Pairs with Local Item Dependence Booklets 7 and 8

Booklet	Passage	Item Pair	Q_3 Value
7	Where's the Honey	14 and 19	-0.248
7	Where's the Honey	19 and 20	0.440
7	Where's the Honey	19 and 21	0.221
8	Flower's on the Roof and Where's the Honey	6 and 20	-0.217
8	Flower's on the Roof and Where's the Honey	8 and 20	-0.229
8	Where's the Honey	20 and 21	0.516
8	Where's the Honey	20 and 23	0.311

As shown in Table 4.16, three of the item pairs with high Q_3 values from 'The Empty Pot' passage in Booklet 3 were also flagged for local dependence in Booklet 12. Two items flagged for local dependence in the passage 'Where's the Honey' from Booklet 8 were also flagged in Booklet 12. The Q_3 values for item pairs (17/ 22) and (17/31) indicate that student's performed better than expected on one item and worse than expected on the other.

Table 4.16

Item Pairs with Local Dependence Booklet 12

Booklet	Passage	Item Pair	Q_3 Value
12	The Empty Pot	17 and 18	0.482
12	The Empty Pot	17 and 19	0.323
12	The Empty Pot	17 and 20	0.335
12	The Empty Pot and Where's the Honey	17 and 22	-0.213
12	Where's the Honey	27 and 28	0.529
12	Where's the Honey	27 and 29	0.276

Booklet	Passage	Item Pair	Q_3 Value
12	Where's the Honey	27 and 30	0.388
12	Where's the Honey	29 and 30	0.251
12	The Empty Pot and Where's the Honey	17 and 31	-0.210

In Booklet 13, there were twenty-six item pairs with local dependence, of which twenty pairs included items 31-34. Items 1 and 13 are both designed to assess the reading comprehension process of examining and evaluating content, language and textual elements. Items 31-33 from the 'The Giant Tooth' passage are presented as a three-part fill in the blank table, as shown in Appendix A. These items are all designed to assess student's ability to interpret and integrate ideas and information. EFA factor loadings on a one-factor model for all the items are comparable with loadings ranging from 0.839 and 0.888, which suggest that these items are capturing similar comprehension processes. For items 31-33 students are asked to describe what a fossil expert thought an Iguanodon looked like in the 1800's compared to what scientists today think it looked like. These three items are closely related. They all ask about scientist's assessments of Iguanodon's physical characteristics. Items 24 and 25 are also a two-part fill in the blank table from 'The Giant Tooth' passage. The Q_3 values for item pairs ranged from 0.20 to 0.83. As shown in Table 4.17, item pairs from different passages resulted in negative LID. The quantity of LID pairs and the large Q_3 values suggest that the standard errors of test scores may be under-predicted in Booklet 13.

Table 4.17

Item Pairs with Local Dependence Booklet 13

Booklet	Passage	Item Pair	Q_3 Value
13	Enemy Pie	1 and 13	0.762
13	Enemy Pie and The Giant Tooth Mystery	1 and 31	-0.325
13	Enemy Pie and The Giant Tooth Mystery	1 and 32	-0.213
13	Enemy Pie and The Giant Tooth Mystery	1 and 34	-0.280
13	Enemy Pie and The Giant Tooth Mystery	7 and 31	-0.300
13	Enemy Pie and The Giant Tooth Mystery	8 and 31	-0.207
13	Enemy Pie and The Giant Tooth Mystery	10 and 31	-0.220
13	Enemy Pie and The Giant Tooth Mystery	10 and 32	-0.214
13	Enemy Pie and The Giant Tooth Mystery	12 and 31	-0.267
13	Enemy Pie and The Giant Tooth Mystery	13 and 31	-0.298
13	Enemy Pie and The Giant Tooth Mystery	13 and 34	-0.275
13	Enemy Pie and The Giant Tooth Mystery	14 and 31	-0.245
13	Enemy Pie and The Giant Tooth Mystery	15 and 31	-0.230
13	The Giant Tooth Mystery	19 and 31	-0.221
13	The Giant Tooth Mystery	23 and 31	-0.218
13	The Giant Tooth Mystery	24 and 31	-0.291
13	The Giant Tooth Mystery	24 and 25	0.833

Booklet	Passage	Item Pair	Q_3 Value
13	The Giant Tooth Mystery	24 and 26	0.699
13	The Giant Tooth Mystery	24 and 26	0.699
13	The Giant Tooth Mystery	24 and 34	-0.273
13	The Giant Tooth Mystery	25 and 31	-0.211
13	The Giant Tooth Mystery	26 and 31	-0.256
13	The Giant Tooth Mystery	26 and 34	-0.258
13	The Giant Tooth Mystery	31 and 32	0.589
13	The Giant Tooth Mystery	31 and 33	0.428
13	The Giant Tooth Mystery	31 and 34	0.598

Identification of DIF using IRT. The IRT based LH method of DIF detection was conducted with PARDUX software (CTB/McGraw-Hill, 1991), which computes the difference between the observed and expected p-values for each item by deciles of the specified group. A Z-statistic is calculated for each decile and an average Z-statistic is computed for items to identify the degree of DIF. Items with a Z-statistic ≥ 2.58 and $|p_{diff}| < 0.10$ are identified as moderate magnitude DIF, Level 2. Large magnitude DIF, Level 3 is identified by $|Z| > 2.58$ and $|p_{diff}| \geq 0.10$. The interpretation of items classified as moderate and large DIF is that parameters for those items are not invariant across the two language groups (Hambleton et al., 1991).

Results of the IRT DIF detection for all thirteen booklets are summarized in Table 4.18. As shown in Table 4.18, a total of six items were identified as either moderate or large DIF in Booklet 1 with five in favor of the English language group and one in favor of the French language group. A total of eight items were identified as DIF in Booklet 2 with four favoring the English language group and four favoring the French language group. In Booklet 3, seven of 34

items were detected as DIF. Two of those items favored the English language group and five favored the French language group. In Booklet 4, eight of 32 items were detected as DIF, with three favoring the English language group and five favoring the French language group. Six out of 24 items were identified as DIF in Booklet 5, with four items favoring the English language group and two favoring the French language group. In Booklet 6, eight of twenty-four items were identified as DIF, with four items favoring each language group. In Booklet 7, six out of 28 items were identified as moderate DIF, with four items favoring the English language group. In Booklet 8, eight out of 29 items were identified as DIF. Five of the items identified in Booklet 8 favored the English language group and three favored the French language group. In Booklet 9, ten out of 25 items were identified as DIF, with five favoring each language group. In Booklet 10, six of twenty-four items were identified as DIF, with four in favor of the English language group and two in favor of the French language group. In Booklet 11, 5 out of 26 items were identified as DIF, with two favoring the English language group and three favoring the French language group. In Booklet 12, four of 36 items were identified as DIF with two favoring each language group. Twelve of thirty-five items were identified as DIF in Booklet 13. Six of those items were identified as DIF in favor of the English language group and six were identified as DIF in favor of the French language group. Of those that favored the English language group, three were identified as very large DIF, with $|p_{diff}| > .35$.

Table 4.18

Number of DIF Items detected by LH IRT for Booklets 1-13

		Pro-English			Pro-French			
Booklet	Total # of Items	Level 2	Level 3	Total # for English	Level 2	Level 3	Total # for French	% of DIF items
1	6	1	4	5	0	1	1	24%

Booklet	Total # of Items	Pro-English			Pro-French			% of DIF items
		Level 2	Level 3	Total # for English	Level 2	Level 3	Total # for French	
2	7	3	1	4	2	1	3	27%
3	7	1	1	2	3	2	5	21%
4	8	1	2	3	4	1	5	25%
5	6	1	3	4	1	1	2	21%
6	8	2	2	4	2	2	4	33%
7	6	4	0	4	2	0	2	21%
8	8	3	2	5	2	1	3	28%
9	10	1	4	5	4	1	5	40%
10	6	1	3	4	1	1	2	25%
11	5	1	1	2	1	2	3	19%
12	4	2	0	2	1	1	2	11%
13	12	3	3	6	5	1	6	34%

Identification of DIF using LR/OLR. The LR method was used to verify DIF detection by the LH IRT method. Results of LR analysis for all thirteen booklets are shown in Table 4.19. In Booklet 1, all the items identified as DIF by LR and OLR corroborate those identified by IRT DIF. In Booklet 2, the LR method detected moderate or large DIF in four of the items identified by the IRT DIF method. In Booklet 3 five of the seven items identified as DIF by the IRT method were identified by the LR method. Four of the eight items identified by IRT DIF were identified as moderate or large DIF by the LR method in Booklet 4. The four additional items identified by the IRT DIF method were identified as negligible DIF by the LR method. In Booklet 5, the LR method identified the same five items as the IRT DIF method. Three of the items identified by the LR method in Booklet 6 verify those identified by the IRT DIF method. An additional five items identified as moderate DIF by the IRT method were identified as

negligible DIF by the LR method. In Booklet 7, one of the six items identified by the IRT DIF method was identified as moderate DIF by the LR method. The additional five items identified by the IRT DIF method were identified as negligible DIF by the LR method. In Booklet 8 four of the eight items identified by the LH IRT method were identified by the LR method. The four additional items identified by the LH IRT method were classified as negligible DIF by the LR method. For Booklet 9, four items were verified by the LR method as moderate or large DIF. The additional seven identified by the IRT DIF method were identified as negligible DIF by the LR method. In Booklet 10, five of the six items identified by the LR method were consistent with the IRT DIF method, with the additional item identified as negligible DIF. Of the five items identified as moderate or large DIF by the LH IRT method in Booklet 11, one was identified as moderate DIF by the LR method. The four additional items were identified as negligible DIF by the LR method. Of the four items identified as DIF by the LH IRT method for Booklet 12, one was identified as large DIF by the LR method, with the additional three items identified by LR as negligible DIF. In Booklet 13, the LR method identified three of the twelve items identified by the IRT method as moderate DIF and nine as negligible DIF. The LR method identified one item as DIF, which was not identified by the LH IRT method.

Table 4.19

Items Identified as DIF by LR and OLR Methods for Booklets 1-13

Booklet	Total # of Items	Moderate DIF	Large DIF	DIF Type	% of DIF Items
1	6	4	2	Uniform	24%
2	4	4	0	Uniform	15%
3	5	3	2	Uniform	15%
4	3	1	2	Uniform	9%

Booklet	Total # of Items	Moderate DIF	Large DIF	DIF Type	% of DIF Items
5	5	2	3	Uniform	21%
6	3	2	1	Uniform	13%
7	1	1	0	Uniform and non-uniform	4%
8	4	3	1	Uniform	14%
9	4	1	3	Uniform and non-uniform	16%
10	5	5	0	Uniform	21%
11	1	1	0	Uniform	4%
12	1	0	1	Uniform	3%
13	4	4	0	Uniform	11%

Table 4.20 compares the number of items identified as DIF by the IRT and LR methods for all the booklets. The LR DIF method detected DIF for all 74 of the items identified by the IRT method; however, only 27 of these items were classified as moderate and large DIF. Differences between these two methods were primarily based on the magnitude of variance between the two language groups, with the LH method detecting greater effect sizes.

Table 4.20

Number of DIF Items Identified by IRT and LR DIF Methods

Booklet			LR	
			DIF	No DIF
1	LH	DIF	6	0
		No DIF	0	18
2	LH	DIF	4	3
		No DIF	0	19
3	LH	DIF	5	2
		No DIF	0	27

Booklet			LR	
			DIF	No DIF
4	LH	DIF	3	5
		No DIF	0	24
5	LH	DIF	5	1
		No DIF	0	18
6	LH	DIF	3	5
		No DIF	0	16
7	LH	DIF	1	5
		No DIF	1	21
8	LH	DIF	4	4
		No DIF	0	21
9	LH	DIF	4	6
		No DIF	0	15
10	LH	DIF	5	1
		No DIF	0	18
11	LH	DIF	1	4
		No DIF	0	21
12	LH	DIF	1	4
		No DIF	0	31
13	LH	DIF	3	9
		No DIF	1	22

It is important to note that DIF methods are known to differ with respect to the items detected and effect size (Ercikan et al., 2004; Kristjansson et al., 2005; Sireci et al., 2003). The differences have to do with how these methods estimate the probability of responding correctly and the matching level criterion. The LH IRT model uses marginal maximum likelihood estimation to estimate the probability of responding correctly for a range of ability levels by jointly estimating item parameters and theta levels. The LR/OLR method estimates the probability of responding correctly by dividing the number of people who obtain a maximum

score by the number of people in the matched group. The matching criterion used by the LH IRT method is ability level based on joint estimates of item parameters and theta levels. With the LR/OLR method, the matching criterion is the number correct or total score. The LR/OLR method does not account for the range of item difficulties and the interaction between ability levels and item parameters. As a result of the differences, these methods differ in their levels of power to detect DIF.

The advantage of the LR method is that it detects uniform and non-uniform DIF. Uniform DIF occurs when there are discrepancies between the two groups with difficulty parameters. With uniform DIF, group differences on an item are the main effect and are not dependent on where an examinee scores on the reading literacy ability continuum. Non-uniform DIF occurs when there are differences between the group responses across the levels of reading literacy ability for an item. Hence, the interaction of group by ability after statistically matching on the test score results in non-uniform DIF.

Summary

Results from the two DIF detection methods indicate that for thirteen of the PIRLS 2011 booklets administered in Canada items from the French and English language versions function differently. For thirteen of the booklets items functioned differently in the range of 11% to 40%, averaging 25%. There were more items in favor of the English language group (50) than the French language group (43). Of the DIF items, 26 were large magnitude in favor of the English language group compared to 15 large magnitudes in favor of the French language group. The probability of responding correctly for French and English language students of equal reading ability are not equivalent for the items identified as DIF from PIRLS 2011 administered in Canada. For DIF items, the probability of responding correctly is based on group membership,

which implies that other factors account for performance differences. Previous research in Canada on educational LSAs has found evidence of DIF across French and English language versions ranging from 14% to 40% (Ercikan, 1998; Ercikan, 2002; Ercikan et al., 2004; Ercikan et al., 2010; Gierl & Khaliq, 2001; Oliveri & Ercikan, 2011). Results from this study support previous findings' demonstrating differences for French and English language versions of LSAs at the item level.

Research Question Three: Blind expert reviews of DIF and non-DIF items

Expert reviewers evaluated the equivalence of the French and English language versions of four passages from PIRLS 2011. The experts were all fluent in both languages and two of the experts have extensive experience with test development and test equivalence across French and English language versions and expertise in reading literacy. Passages were reviewed in the sequence of the booklets. They first reviewed the passage 'Fly, Eagle, Fly' from Booklets 1, 2 and 10. Next they reviewed the passage 'Day Hiking' from Booklets 5, 6 and 10, followed by 'Enemy Pie' and 'The Giant Tooth Mystery' from Reader Booklet 13. Expert reviewers reviewed all of the items that were identified as DIF by statistical methods and an equal number of randomly selected non-DIF items. French-English bilingual reviewers were instructed to focus specifically on language, cultural and format differences and to judge whether the differences were expected to result in performance differences. The results are organized according to the types of differences noted for items between language versions within each passage. Text differences noted by reviewers are also discussed by passage. Items given a rating of 2 or 3 in the 'Fly, Eagle, Fly' passage are displayed in Table 4.21.

Table 4.21.

Passage 1 ‘Fly, Eagle, Fly’ Expert Review Ratings and Noted Differences

Item	Rating	Favors French or English	Noted Differences
1	2	French	Differences in verb tense, in words, expressions and structure of sentence inherent to a language or culture and differences in length and complexity of item. The English version asks, “What did the farmer set out to look for?” The French version asks, “What is he looking for?”
2	2	Unclear which group it favors	Differences in verb tense.
3	2	Unclear which group it favors	Differences in cohesiveness and continuity of text and in additional information that may guide students thinking. In English, the question asks, “What in the story shows…” while in French is asks, "qu'est-ce qui montre…" The word "montre" is a literal translation but the word "démontre" would be more appropriate. In English, a response option describes bringing the eagle chick to his family, while in French the option describes bringing the eagle chick home.
7	2	Unclear which group it favors	Differences in verb tense and in words, expressions and structure of sentence inherent to a language or culture. The French version seems to suggest where one’s place is while the English version suggests a sense of belonging to a place. The English version states, “you belong not to the earth but the sky.” The French version states, “Ta place n'est pas sur terre mais dans les airs.”

Item	Rating	Favors French or English	Noted Differences
8	2	Unclear which group it favors	Differences in verb tense and in words, expressions and structure of sentence. The English version uses past tense and the French version uses present tense.
11	2	Unclear which group it favors	Differences in verb tense and in words, expressions and structure of sentence. The English version uses past tense and the French version uses present tense. The English version asks, "Why was the rising sun important to the story?' The French version asks, "Pourquoi le lever du soleil joue-t-il un role si important dans l'histoire?"
12	2	English	Differences in words, expressions and structure of sentence inherent to a language or culture. In the English version the phrase "things that he did" was translated to "comportement." Differences in word difficulty or familiarity of vocabulary. Differences in additional information that guides how examinees think. The English version says, "Describe what the friend was like." The French version says, "Décris son caractère." The English version specified the friend but the French version does not.

In passage 1 reviewers identified differences for most of the items they examined; however, only eight items were classified as moderate or large magnitude. Three of the items they identified as having moderate or large magnitude differences were not statistically identified as DIF. For three of the four items identified by expert reviewers in this passage, reviewers were unclear which group the differences favored, in part because they identified multiple differences that alternately favored one or the other group. Reviewers noted that in this passage there were differences in word difficulty and familiarity of vocabulary between the two language tests, and

inconsistencies with verb tense, wording and sentence structure throughout the text. For instance, a number of words that were used in the French version are uncommon for Francophone lexicon. Words such as, 'vachers', 'montrent' and 'aboyer' were identified as unfamiliar. The phrase "things that he did" in item 12 was translated as 'comportement' in the French language version. Verb tenses differed between language versions, with the English version written in past tense and the French version written in present tense. Reviewers specified that differences with wording and sentence structure in this test might cause language groups to offer slightly different answers.

Of the four passages that were examined, reviewers identified passage 2 'Day Hiking' as having the greatest number of differences between the two language versions. Of the nine items that were examined from this passage, reviewers identified differences for every item, with eight classified as moderate or large magnitude. Three of the items classified as having moderate or large magnitude differences by reviewers were not statistically identified as DIF. The passage 'Day Hiking' in English was titled "Découvre les joies de la randonnée" in French. Expert reviewers identified numerous differences in passage 2 that made the French version confusing and longer than the English version. They attributed some of the vocabulary differences in this passage to the use of coined expressions in the English version that did not easily translate to French. There were also a number of French words that reviewers described as less common to French language groups in Canada. They attributed many of these differences to inappropriate translation. The font size was noticeably smaller in the French language version and there were a number of punctuation and layout differences between the two language versions. The table in the French version contained an English title and English subtitles. The word count for the English version was 1200 while the word count for the French version was 1400. Several of the

items in the English language version were worded in the form of a question and worded as a statement in the French language version. There was also consensus among the reviewers that key words and phrases varied between text and questions for the French version. For example, the word hiking was translated to ‘randonnée’ but then it was switched to the word ‘plein air’. Reviewers agreed that the word ‘plein air’ has a different meaning than the word ‘hiking’. In the English version hiking was tied to the subject of the entire book while in the French version several items do not mention going on a hike. For instance, item 8 explains that you are returning from your hike while the French version is stated as when you return, without mention of a hike. Reviewers identified a number of inconsistencies within the text and between the text and questions in the French version. For instance, the English version used the term ‘map key’ to refer to a legend, while the French version used several terms including ‘légend’ and ‘tableau qui accompagne la carte’. All items rated level 2 or 3 in passage 2 by reviewers are displayed in Table 4.22. For items that were not discussed as a group in the second stage of the review process, ratings represent the average of the four reviewer ratings.

Table 4.22

Passage 2 ‘Day Hiking’ Expert Review Ratings and Noted Differences

Item	Rating	Favors French or English	Noted Differences
1	3	French	Differences in words, expressions, and structure of sentence. Differences in additional information that guides student’s thinking. Differences in reading processes assessed. In the French version, a key term switches from "randonnée" to "plein air." The English version asks for student’s impression by asking, “What is the main message?” The French version asks, “What is the main idea?”

Item	Rating	Favors French or English	Noted Differences
2	2	English	Differences in additional information that guides how examinees' think. The English version cues the reader to search in the leaflet in the beginning of the sentence ("the leaflet said…"), while it is in the last part of the question in the French version ("d'apres le dépliant").
3	3	English	Differences in cohesiveness and continuity of text. Differences in words, expressions, and structure of sentence. The item is written as a question in English and as a statement in French. Differences in meaning. The English version asks, "What are the two things the leaflet told you to keep in mind?" The French version asks, "Nomme deux points, décrits dans le dépliant." Differences due to inappropriate translation. In the English version you *are* hiking while in the French version you *leave* for a hike.
5	2	English	Differences in words, expressions and structure of sentence. Differences in word difficulty and or familiarity of vocabulary. The word 'blisters' is used in the English version and the word 'ampoules' in the French version. Differences in meaning.
6	2	English	Differences in word difficulty or familiarity of vocabulary. In English, the question asks, "What should you do if you get into trouble while you are hiking?" The English version sounds more urgent, while in French, it suggests general difficulties, but the urgency is ambiguous. In French, the question asks, "Que dois-tu faire si tu as des problèmes pendant ta randonnée?" Many of the answer choices in the French version are acceptable solutions to 'si tu as des problèmes.'

Item	Rating	Favors French or English	Noted Differences
7	2	English	Differences in word difficulty or familiarity of vocabulary. Differences in length or sentence complexity that make the item more difficult for one language group. The phrase, "si tu as des problèmes" is not equivalent to "if you get into trouble". The reading load for the instructions to this question is much higher in French.
8	3	English	Omissions or additions of worlds or phrases in one language version that affect meaning. The English version states that you are returning from your hike while the French version simply states when you return. Also the item in the French version is not worded in a question form as it is in the English version.
9	2	English	Differences in words, expressions and structure of text and table. Differences in length and sentence complexity. The French version is considerably longer and the wording is awkward and difficult. In the French version the terms ‘map key’ and ‘legend’ vary between the text and question. The English version asks what Tom was surprised about in the day, while the French version asks what surprised Tom throughout the day. The table in the French version contains an English title and English subtitles. Names of destinations differ. The name of a destination in English is ‘Lookout Hill Circle’ and in French it is ‘Randonnée autour de la colline du Guet.’
11	3	Unclear which group it favors	Differences in word difficulty or familiarity of vocabulary. Differences in length or sentence complexity. The term ‘map key’ is used in English while the phrase ‘tableau qui accompagne la carte’ is used in the French question. In the French text the word ‘légend’ is used but then the phrase ‘tableau

Item	Rating	Favors French or English	Noted Differences
			qui accompagne la carte' is used for the question. French version is longer but easier to understand. The English version requires reader to 'study' the key while the French version requires the reader to 'observe' the key.

Reviewers agreed that in passage 3 titled 'Enemy Pie' in English and 'La tarte des ennemis' in French there were many differences with word difficulty or familiarity of vocabulary, as well as differences in the choice of expressions and structures of sentences that made the text and questions in the French version more complex. For instance, the word 'feel' is translated as 'réagi'. Item 3 in the English version states, "Write one ingredient that Tom thought would be in Enemy Pie." In the French version item 3 states, "Écis un des ingrédients que Thomas s'attendait à trouver dans la tarte des ennemis." Overall, with the exception of one item reviewers identified the differences as being in favor of English language students in this passage. As with the first two passages, they attributed many of the differences to poor and inappropriate translation. Reviewers identified three items as having moderate or large magnitude differences that were not identified statistically as DIF. Additional examples of the types of differences identified in passage 3 are shown in Table 4.23.

Table 4.23

Passage 3 'Enemy Pie' Expert Review Ratings and Noted Differences

Item	Rating	Favors French or English	Noted Differences
6	2	English	Differences in words, expressions and structure of sentence inherent to a language or culture. Differences in word difficulty or familiarity of vocabulary. For the English

Item	Rating	Favors French or English	Noted Differences
			version it says, “Write one thing.” For the French version it says, “Écris une conséquence.”
7	2	English	Differences in meaning. Differences due to inappropriate translation. The English versions asks, “What were the two things…” while the French version asks, “Le père a fait deux recommandations…” The word thing is translated as recommendations.
9	3	English	Differences in meaning. Differences in length or sentence complexity that make the item more difficult for one language group. The English version asks, “What surprised Tom *about* the day?” The French asks, “What surprised Tom *during* the day?” The translated French version of this question is also awkward.
10	2	Unclear which group it favors	Differences in meaning. Differences due to inappropriate translation. The English version asks why Tom should forget about the pie, while the French version asks why Tom should avoid the pie. The phrase ‘at dinner’ was translated as ‘au repas,’ which is not the same. The phrase ‘piece of enemy pie’ is translated as ‘la part de la tarte d'ennemi.’
13	3	English	Differences in words, expressions and structure of sentence. Differences in length or sentence complexity that make the item more difficult for one language group. The English version asks, “What does this suggest about the boys?” The French version asks, “Qu'est-ce que cette phrase permet de conclure?”
15	2	English	Differences in meaning. The English version asks, “What kind of person is Tom’s dad?” The French version asks, “Quel genre de

Item	Rating	Favors	Noted Differences
		French or English	
			personne est le père Thomas?" Genre also means gender. The obvious answer is to look for a masculine reference in the text.

In passage 4, as shown in Table 4.24 titled "Giant Tooth" in English and "Le mystère de la dent Géante" there was a general agreement among reviewers that there were differences in words, expressions and structure of sentences inherent to Francophone language and culture that made the text in the French version more difficult than the English version. For example the phrase 'looked like' was translated as 'apparence extérieure.' They also noted differences in word difficulty or familiarity of vocabulary in the French text with words such as 'caractèr' for the word 'gulped'. They questioned the use of the word 'piquant' for the word 'spike'. Overall, reviewers decided that the differences between the two language versions for this passage were largely due to poor and inappropriate translation, which added an additional layer of complexity to questions in the French language version. Item 15 was identified as an example of inappropriate translation. The English language version states, "later discoveries proved that…" while the French version states, "les découvertes suivantes ont prouvé…" Reviewers commented that a more appropriate translation would have been "les découvertes subséquentes ont prouvé…" For this passage, one item was classified as having moderate or large magnitude differences that was not statistically identified as DIF.

Table 4.24

Passage 4 ‘The Giant Tooth Mystery’ Expert Review Ratings and Noted Differences

Item	Rating	Favors English or French	Noted Differences
2	2	English	Differences in words, expressions and structure of sentence inherent to a language or culture. The English version uses the phrase “long ago” and the French version uses the phrase “Il y a très longtemps.”
7	3	English	Differences in word difficulty or familiarity of vocabulary. Differences in length or sentence complexity that make the item more difficult for one language group. The English version asks, “What did Gideon Mantell know about reptiles that made the fossil tooth puzzling?” The French version asks, “Que savait Gideon Mantell sur les reptiles qui lui fait comprendre le caractère intriguant de la dent fossile?” The phrase “lui a fait comprendre” adds another layer of complexity to the question.
13	2	English	Differences in length or sentence complexity that make the item more difficult. Differences in word difficulty or familiarity of vocabulary. Omissions or additions of words or phrases that affect meaning. The English version states, “What Gideon Mantell thought the Iguanodon looked like.” The French version states, “L’apparence extérieure de l’Iguanodon d’après Gideon Mantell à cette époque-là.” The French translation is more complicated and refers to the exterior appearance, which the English version omits.
15	3	English	Differences in additional information that guides how examinees’ think. Differences dues to inappropriate translation. The English version says that the Iguanodon was *over* 30 metres long, while the French

Item	Rating	Favors English or French	Noted Differences
			version says that it measured 30 m (mesurait 30 m). The English version states, "later discoveries proved" and the French version states, "Les découvertes suivantes ont prouvé." Reviewers suggested the translation "Les découvertes subséquentes ont prouvé."

Summary

In summary, expert reviewers identified a great number of differences between the English and French versions for all four passages. Sources of differences identified by reviewers were largely due to length or sentence complexity, word difficulty or familiarity with vocabulary, differences in meaning and differences due to inappropriate translation. In general, they found that differences favored the English language group but for several items in passages 1 and 2 reviewers were unsure if the differences favored one linguistic group over another.

Correspondence between DIF identification and expert reviews. In the passage 'Fly, Eagle, Fly' expert reviewers identified four of the items that were detected as DIF by statistical methods. Item 2 in the passage was identified as moderate DIF in favor of the French language group by the IRT DIF method and as negligible DIF by the LR method. Reviewers identified this item as moderate DIF, citing differences in verb tense between the language versions. Item 3 was identified as large DIF in favor of the English language group by both statistical methods and by expert reviewers. As detailed in Table 4.24, reviewers noted differences in verb tense and in additional information that guided examinees thinking with this item. The IRT and LR DIF methods identified item 8 as moderate DIF in favor of the English language group. Reviewers identified this item as moderate DIF, citing verb tense differences between the two language

versions. Item 12 was identified as large DIF by the LH IRT method in favor of the English language group. Reviewers identified this item as moderate DIF in favor of the English language group. In this passage the reviewers attributed differences between the two language versions to differences in words, expressions and structure of sentence that are inherent to a language and culture, differences in word difficulty, differences in additional information that guided examinees thinking and differences in reading processes assessed by the two language versions. They attributed many of these differences to inappropriate translation and the use of words that are uncommon to the Francophone language. For example, one reviewer suggested using the word 'japper' rather than the word 'aboyer' for 'bark' in the English language version. Expert reviewers did not identify item 2, which was statistically detected as DIF.

Consistency between expert ratings and statistical analysis for all the moderate or large magnitude DIF items that were reviewed are shown in Table 4.25. Items classified as having either moderate or large magnitude differences by reviewers that were not statistically identified as DIF are also displayed in the last column of Table 4.25. The sources of differences for each item are displayed in Tables 4.21-4.24.

Table 4.25

Consistency Between Expert Reviews and Statistical Methods by Passage

	IRT	LR	Expert reviews of DIF items	Expert reviews non-DIF items
Item by passage	(favors)	(type)	(favors)	(favors)
1 EF				moderate (French)
2 EF	3.215 (French)		moderate (unclear)	
3 EF	4.641*	53.70*	moderate	

Item by passage	IRT (favors)	LR (type)	Expert reviews of DIF items (favors)	Expert reviews non-DIF items (favors)
	(English)	(uniform)	(unclear)	
5 EF	4.765* (French)	59.27 (uniform)		
7 EF				moderate (unclear)
8 EF	4.425 (English)	37.03 (uniform)	moderate (unclear)	
11 EF				moderate (unclear)
12 EF	3.215 (English)		moderate (English)	
2 DH	3.352 (French)		moderate (English)	
3 DH	4.182* (English)	39.68 (uniform)	large (English)	
5 DH				moderate (English)
6 DH				moderate (Engish)
7 DH	5.027* (French)	67.42* (uniform)	moderate (English)	
8 DH				large (English)
9 DH	2.609 (English)	26.76 (uniform)	moderate (English)	

Item by passage	IRT (favors)	LR (type)	Expert reviews of DIF items (favors)	Expert reviews non-DIF items (favors)
11 DH	3.220* (English)	57.33 (uniform)	large (unclear)	
1 EP	20.277* (English)	1205.63* (uniform)		
6 EP				moderate (English)
7 EP		53.41 (both)		moderate (English)
9 EP				large (English)
10 EP				moderate (unclear)
12 EP	2.636 (English)			
13 EP	20.209* (English)	1253.64* (uniform)	large (English)	large (English)
15 EP	3.627* (English)		moderate (English)	moderate (English)
2 GT	3.817 (English)		moderate (English)	
6 GT	3.452 (French)			
7 GT	4.478 (English)	54.41 (both)	large (English)	
11 GT	2.852 (French)			

Item by passage	IRT (favors)	LR (type)	Expert reviews of DIF items (favors)	Expert reviews non-DIF items (favors)
12 GT	2.656 (French)			
13 GT				moderate English
15 GT	5.786* (French)		large (English)	
16 GT	3.610 (French)			
18 GT	5.385 (French)			

Note. EF refers to the Fly, Eagle, Fly passage. DH refers to the Day Hiking passage. EP refers to the Enemy Pie passage and GT refers to The Giant Tooth passage. Items identified as large DIF are denoted by *

For passage 2 titled 'Day Hiking' expert reviewers identified all five of the items identified as DIF by both the IRT and LR methods. Reviewers documented six types or sources of differences for the five items identified as DIF by statistical methods. These included differences in words, expressions and structure of sentences inherent to the languages, differences in word difficulty or familiarity of vocabulary, differences in meaning, differences in length or sentence complexity that made the French items more difficult, differences due to inappropriate translation and differences with the tables in the two versions.

Consistency between expert review ratings and statistical methods varied for the passage entitled 'Enemy Pie.' Of the four items identified as DIF by statistical methods, reviewers identified clear differences for two of those items. Reviewers also identified clear differences between language versions for an additional four items not identified by statistical methods, although one item was identified by the LR method. Item 1 was identified as large DIF by

statistical methods but expert reviewers did not report clear differences between the two language versions. Item 7, which was identified as moderate DIF by the LR method, was not detected by the IRT method but was rated as having moderate differences by the expert reviewers. Item 12 was identified as moderate DIF by statistical methods but only one of the four reviewers identified moderate differences between the two language versions citing differences in word difficulty or familiarity of vocabulary. Reviewers identified Item 13 as having clear differences that were likely to lead to performance differences between the two language groups. Statistical methods detected Item 15 as large DIF and reviewers identified it as moderately different in favor of the English language group. They cited differences in meaning between the two versions. In the English version the question was, "What kind of person is Tom's dad?" In French the question was, "Quel genre de personne est le père Thomas?" Reviewers also noted that the phrase "pour expliquer ta réponse" in the French version is not the same as "that shows this" in the English version. Overall, with the exception of one item reviewers identified the differences as being in favor of English language students in this passage. As with the first two passages, they attributed many of the differences to poor and inappropriate translation.

In the passage titled 'The Giant Tooth' there was consensus between the four reviewers on four of the items identified as DIF by the two statistical methods. Differences for two of the items identified as DIF by statistical methods and the reviewers were due to omissions or additions of words or phrases that affect meaning and additional information that guides how examinees think. For instance, the English version for item 13 states, "What Gideon Mantell thought the Iguanodon looked like." The French version states, "L'apparence extérieure de l'Iguanodon d'après Gideon Mantell à cette époque-là." The French translation is more complicated and refers to the exterior appearance, which the English version omits. Item 2 was

rated as moderately different and reviewers noted that the item was awkward in both languages. The phrase “Il y a très longtemps” in the French version was identified as placed oddly in the question. There were discrepancies between raters on seven of the items that were statistically identified as DIF. Ratings differed with respect to the degree of differences detected by reviewers on several items. For instance, item 14 was identified as moderate DIF by statistical methods and moderately different by one reviewer, while two reviewers rated differences as minor. The phrase “apparence extérieure” in the French version was cited as clearly different from the phrase “looked like” in the English version. The reviewer identified several sources of differences between the French and English language versions of item 14 including, differences in words, expressions and structure of sentence inherent to a language, differences in word difficulty and familiarity of vocabulary and differences in length or sentence complexity making the French language version more complex. As shown in Table 4.24, reviewers confirmed the moderate and large DIF status of items 13 and 15 in this passage. However, reviewers identified item 13 as favoring the English language group but this item was identified as favoring the French language group by the IRT DIF method. The French phrase “quand il a essayé d'imaginer l'apparence extérieure d'un Ig” was evaluated as more complicated than the English phrase, “when trying to figure out what the Ig looked like.” Item 15 was identified as an example of inappropriate translation. The English language version states, “later discoveries proved that…” while the French version states, “les découvertes suivantes ont prouvé…” Reviewers commented that a more appropriate translation would have been “les découvertes subséquentes ont prouvé…”Overall, reviewers agreed that the differences between the two language versions for this passage were largely due to inappropriate translation, which added an additional layer of complexity to questions in the French language version.

Summary

Expert reviewers identified several sources of differences for items detected as DIF by statistical methods. Although reviewers were clear that there were differences between the two language versions, they were not always clear which group the differences favored. Most sources were attributed to differences in words, expressions and sentence structure inherent to the French language, differences in word difficulty and or familiarity of vocabulary, differences in length and complexity of sentences, differences in meaning and differences due to inappropriate translation. Reviewers attributed a majority of these differences to poor and inappropriate translation. As previously mentioned, reviewers also identified an average of three additional items as having moderate or large differences that were not statistically identified, with a majority of the items favoring the English language group. A variety of DIF detection methods are recommended to corroborate findings because variance across methods is expected. Expert reviews are primarily employed to identify potential sources of differences for items statistically identified as DIF and for examining linguistic and cultural equivalence with respect to meaning, cultural relevance and difficulty of cognitive requirements based on language. Expert reviews are used in combination with statistical methods to provide support for comparability of different language versions. Previous research has shown that although there is a considerable amount of agreement in the identification of DIF between statistical methods and expert reviews, experts do not consistently identify and distinguish DIF and non-DIF items (Ercikan, K., & Lyons-Thomas, J. 2013).

Chapter V: Discussion

Degree of Test Equivalence

The purpose of this study was to examine the degree of measurement equivalence between the French and English language versions of PIRLS 2011 administered in Canada. Statistical analysis and expert reviews were used to evaluate the extent of measurement equivalence and the comparability of scores at the item and test level for French and English language groups. To make comparative inferences across languages and cultures on the basis of test scores requires evidence of item and test level equivalence to ensure comparability of test scores. Evidence must demonstrate that the same construct is being measured at a comparable level of difficulty and that test scores are on a common scale. In this study, CFA results indicate that there are differences in the relation of the items to a one-factor model for thirteen booklets across the two language groups. A one-factor model demonstrated a poor statistical fit for all thirteen booklets, based on the Chi-square criteria and a poor statistical fit for ten booklets based on multiple criteria. Therefore, a one- factor model does not represent the data in the same way for the French and English language versions of PIRLS 2011 across thirteen booklets. The CFA results can be interpreted to mean either that the same factor structure does not hold for the two language groups or that a one factor model is mis-specified in one or both groups (Meade & Kroustalis, 2006). Further evidence from EFA results indicated that although one dominant factor accounted for the largest proportion of variance across the thirteen booklets, the number of potential factors were not similar for the two language groups across ten of the booklets. In addition, different patterns of loadings on factors explain the variance-covariance matrices across ten of the booklets for the two language groups. Although this evidence is not conclusive, it does

suggest that the differences are significant enough to warrant further examination, which is discussed in the future research section of this chapter.

An examination of item level equivalence using two DIF detection methods demonstrated that an average of 25% of items across all thirteen booklets function differently across language versions. Fifty items were identified as DIF in favor of the English language group across the thirteen booklets, with 26 of those items classified as large magnitude DIF. Forty four items were identified as DIF in favor of the French language group across the thirteen booklets, with 15 of those flagged as large magnitude DIF. These results demonstrate a lack of item equivalence across the two language versions.

Results from expert reviews of the four passages indicate that there are many differences between the two language versions, due primarily to linguistic and cultural differences. Results from expert reviews imply that the adaptation process produced unintended differences in content and difficulty levels between the two language versions.

As a whole, evidence from this study indicates there are important scale level differences between the French and English language versions of PIRLS 2011 that call for further investigation and there is a lack of equivalence at the item level.

Implications

The results of this study have implications at several phases of testing practices. These include the test adaptation procedures that are used to establish and ensure equivalence, the validity of inferences based on score comparability of PIRLS 2011 for French and English language groups and methods for effectively assessing test equivalence. These implications are described below.

The first implication based on results from this study is that test adaptation procedures for PIRLS in Canada may need to be reexamined to determine if more rigorous and standardized procedures should be adopted. This implication is based on a) evidence from previous research that test adaptation creates sources of bias in Canada; b) evidence from this study of DIF ranging from 11% to 40% across 13 PIRLS booklets and c) expert review results from this study that indicate many of the differences between language versions for the passages they examined were due to inappropriate translation. *Standard 9.7* in the *Standards* (2012) recommends that test developers describe the procedures used to establish and ensure adequacy of adaptation and provide evidence of score reliability and validity for linguistic groups. PIRLS International Study Center (TIMMS & PIRLS International Study Center, 2012) provides guidelines to countries that participate in PIRLS but each country is responsible for ensuring the appropriateness and quality of the translation. Although translated versions for each country undergo two rounds of verification reviews by linguistic and assessment experts at the international test center, translation procedures are at the discretion of national test centers. Research demonstrates that guidelines are insufficient to ensure high-quality adaptation (Arffman, 2010; Solano-Flores, 2009). A review of current practices for adapting PIRLS in Canada could provide information about the strengths and weaknesses of existing practices and indicate how to create systematic approaches to ensure test equity for French and English language groups.

Cross-cultural measurement researchers have made a number of recommendations for standardizing test adaptation practices that may have applicability in Canada. For instance, the Test Translation Error (TTE) framework Solano-Flores, Contreras-Niño and Backhoff (2005) developed recommends using a team of reviewers with a variety of expertise to address multidimensionality. The TTE framework makes the recommendation to use reviewers with a

variety of expertise because evidence indicates that translation errors are interrelated and multidimensional. Test item properties such as content, language, format, conventions and linguistic demands are interrelated. The interrelations between item properties create tensions for how translators resolve errors. Translators use strategies to resolve tensions that may produce errors in other dimensions. Arffman (2010) found that differences between language versions were related to strategies and choices made by translators. For instance, translators use interference as a strategy in an effort to follow the original text too closely or to improve upon the target text. This strategy resulted in lexical errors, imprecise language, ungrammaticalities and unintelligible language in the Finnish version of PIRLS 2000 (Arffman, 2010). Arffman concluded that it is possible to improve expert reviews through the use of a more standardized procedure. For instance, national test centers can set criteria requiring that qualifications of reviewers extend beyond expertise in the domain assessed to include experience with test development, knowledge of reading processes, response processes, and factors affecting item difficulty (Arffman, 2012). At present, guidelines set out by PIRLS International Study Center recommend that national centers use one skilled and experienced translator, which is insufficient according to research evidence (Arffman, 2013).

The second implication is that test equivalence and score comparability cannot be assumed when tests are adapted for French language groups in Canada. With the increased use of LSAs, it is imperative that organizations such as the International Association for the Evaluation of Educational Achievement and national testing centers provide evidence of score comparability to increase the likelihood that inferences, decisions and consequences are fair, justified and effective. Due to the growing influence of LSAs on curriculum, instruction, school accountability, performance standards and educational policy, the Council of Ministers of

Education must ensure that actions and consequences based on test scores be justified by evidence. Evidence of score comparability is important when CMEC issues reports that the average scores of students enrolled in French language schools are significantly lower than those enrolled in English language schools. Such evidence is particularly important for conclusions such as the one below made by CMEC, "overall, there is a clear pattern in the difference in reading results between students enrolled in the English-language school systems and those in the French-language school systems" (2012, pg. 71). However, it is important to note that the degree of test equivalence and score comparability across French and English groups is likely to vary according to linguistic differences within French language groups (Ercikan, Roth, Simon, Sandilands and Lyons-Thomas, 2013). The degree of incomparability cannot be generalized across all French language groups in Canada. For instance, linguistic differences between French and English language groups may not be as extensive for French language students living in a majority setting. The linguistic background of students may differently disadvantage those living in minority settings and those who do not speak French at home. It is recommended to consider findings from this study in the context of the abovementioned language related factors.

As PIRLS is the only international LSA administered in Canada to assess reading literacy in the early years of education and French and English are the two official languages, the CMEC should test measurement invariance and take reasonable steps to ensure that linguistic groups are given the opportunity to perceive and respond to tests in the same way.

The third implication is that results from this study suggest that measurement incomparability between English and French language groups for PIRLS 2011 accounts for a significant proportion of the observed performance differences. At the scale and item level, score comparability was threatened. The lack of configural invariance for ten of the booklets suggests

that English and French language students may respond to the reading literacy test items differently. The CFA, EFA and DIF results indicate that PIRLS 2011 reading literacy test does not have the same psychometric properties at the scale and item level for the two language groups; therefore, it is likely that observed score differences are an artifact of measurement practices. Expert reviewers indicated that observed score differences were likely due to inappropriate translation. Overall, the differences reviewers found related to words, expressions, meaning, familiarity of vocabulary words, and sentence complexity made the French items more difficult. Results from this study suggest that caution should be exercised with PIRLS 2011 score comparisons between English and French language groups in Canada.

The final implication is that more research is needed to verify the best methods to examine noninvariance and partial noninvariance when working with LSA matrix sampling designs. This implication relates to Messick's (1995) unified validity framework. Messick argues that validity is an evaluative judgment of the degree to which evidence supports the uses, interpretations and consequences of a test. Although recommendations for examining test equivalence are available in the literature, there is a lack of evidence regarding the best methods to examine noninvariance findings at the scale level when conducting simultaneous multigroup CFA. The methods used in this study allowed for verification of results and elucidation for some of the test level differences between the two language versions. Yet, inconclusive test level differences raise questions about the best way to proceed when results indicate noninvariance for a majority of the booklets but not all the booklets. This topic is addressed below in the future directions section of the chapter.

Limitations

There were several limitations to this study. The first is that this study did not include enough information about the diversity within English and French language samples. Students who attend French-language schools but live in English-speaking environments may not have the same exposure to French language outside of school as those living in a French dominant setting such as Quebec where the official language is French by law (Ercikan et al., 2013). Furthermore, the proportion of Canadians that speak either or both of the official languages in Canada varies greatly across provinces. In minority language settings, the composition of French language schools differs substantially. For instance, in Ontario, 55% of the elementary students that attended French language schools in 2006 had one French-speaking parent. A number of students attending French-language schools in Ontario immigrated to Canada from African countries such as Somalia, Ethiopia and Rwanda (Farmer, 2008). Although the focus of this study was not on heterogeneity within language groups, such information highlights the linguistic, cultural, educational and socioeconomic diversities within language groups that affect student performance.

It is also important to note that measurement comparability may look different for different subgroups within language categories. Recent research suggests that linguistic differences may not be as extensive between French and English language groups for French language students living in a majority setting (Ercikan et al., 2013). A recent study examined the accuracy of measurement comparability for French language students living in Quebec versus French language students who live in minority language settings (Ercikan et al., 2013). Comparisons were made between French language students living in majority and minority settings and between students living in minority settings that do speak French at home and those

that do not speak French at home. Differences were found across the three groups in the numbers of DIF items and the items identified as DIF, with larger numbers of DIF items in the comparisons between those living in majority settings and those living in minority settings who do not speak French at home. Results indicated higher reading literacy performance levels for Quebec French Francophone students than for French language students living in minority language settings. Of the three groups compared in this study, French language competency was lowest for students attending French language schools living in minority settings who do not speak French at home. Results from this study provide evidence to substantiate measurement incomparability across French and English test versions for adapted LSAs in Canada. However, results from this study do not address potential differences in measurement comparability within language groups across Canada. Both the number of items and the items identified as DIF are likely to vary for students living in majority and minority settings and for students who speak the test language at home.

The second limitation of this study is related to the expert review process. Although, a two-stage review process was used, the time allotted to the group discussion in the second stage was insufficient. Reviewers provided detailed and extensive information in the first stage when they reviewed the passages and items individually, but in the group discussion when they addressed rating differences their analysis were clearer and more thorough. Although, consensus was reached for all the items discussed as a group, there was not enough time to review rating discrepancies for every item.

Contribution of Findings to Literature

This study contributes to the literature on test equivalence and score comparability across linguistic and cultural groups in several ways. To begin with, it adds to and supports previous

research on the lack of score comparability across French and English language groups of LSAs administered in Canada. Given the increasing demand for accountability in education and the rise of LSAs in Canada, there is an unquestionable need to implement practices and processes in order to ensure the validity of test interpretations across language groups. Reading literacy is fundamental to cultural, political, social and economic growth of a society (CMEC, 2012) and PIRLS is the only international program that assesses reading achievement. If test score results from PIRLS continue to be used to track early literacy skills, to evaluate and remedy educational systems and inform resource allocation decisions in Canada, then it is important to ensure that testing practices are sufficiently rigorous to validate inferences. Expert review results from this study of the four released passages indicate that improvements to the translation process in Canada may reduce some of the sources of differences between the French and English language versions of PIRLS. Finally, results from this study point to the importance of using a variety of methods to investigate test level discrepancies and to the need to gather evidence about the most effective methods for analyzing noninvariance and partial noninvariance.

Future Directions

To elucidate test level differences between the French and English language versions of PIRLS 2011 there are additional statistical analysis that can be conducted. For instance, separate CFA analysis by language for each booklet could clarify the reasons that a one-factor model resulted in a poor fit. Differences may be due to factor structures differences or to model mis-specification for one or both language groups. EFA results suggest that a two, three or four factor models represent the data best for most of the booklets. Simultaneous CFA can be conducted to test a variety of factor models. For instance, after identifying items that operate differently across groups models constraining those items could be tested. Raju, Laffitte and Byrne (2002)

recommend conducting item equivalence tests by constraining an identified item to be equal across the two groups to allow each identified item to be tested.

Further evaluation of item level differences could also include the use of think aloud protocols (TAPs) as an approach to examine and confirm sources of DIF across French and English language versions of tests. With the TAPs approach, examinees are instructed to verbalize their thoughts and understandings of questions as they read through and respond to test questions. Sources of differences between language versions can be examined by using released passages and items to conduct TAPs with both expert reviewers and with students from French and English language groups. In this way, data regarding student's understandings of items, the reasoning used to select or construct answers and the aspects of items that help or impede the ability to problem solve can be collected through the use of TAPs.

As previously mentioned, recent research suggests that students who speak French at home and students who do not speak French at home should be examined separately to better understand performance differences within language groups (Ercikan et al., 2013). Ercikan et al. 2013 recommend conducting controlled studies within French language groups to examine differences between students who live in majority and minority settings and between students who do and do not speak the test language at home. Further research is necessary to examine the extent of test inequivalence within French language groups in Canada.

References

Alexander, P. A., & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. *Handbook of reading research*, 3, 285-310.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of educational measurement*, 36(3), 185-198.

Allalouf, A., & Sireci, S. G. (1998, April). *Detecting sources of DIF in translated verbal items*. Paper presented at the meeting of American Educational Research Association, San Diego, CA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Arffman, I. (2007). *The problem of equivalence in translating texts in international reading literacy studies: a text analytic study of three English and Finnish texts used in the PISA 2000 reading texts*. Institute for Educational Research, Report 21.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research*, *54*(1), 37-59.

Arffman, I. (2012). Unwanted literal translation: An under-discussed problem in international achievement studies. *Education Research International*, *2012*.

Arffman, I. (2012). International education studies: Increasing their linguistic comparability by developing judgmental reviews. *ISRN Education*, vol. *2012*, Article ID 179824.

Arffman, I. (2013). Problems and Issues in translating international educational achievement tests. *Educational Measurement: Issues and Practice*, *32*(2), 2-14.

Arim, R. G., & Ercikan, K. (2005). Comparability between the US and Turkish versions of the Third International Mathematics and Science study's mathematics test results. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1) 1-42.

Bachman, L.F. (2007). Language assessment: Opportunities and challenges. In meeting of the American Association of Applied Linguistics (AAAL), Costa Mesa, CA.

Baker, C. (1992). *Attitudes & Language.* U.K., U.S.A., Multilingual Matters Ltd.

Beller, M., Gafni, N., & Hanani, P. (2005). Constructing, adapting, and validating admissions tests in multiple languages: The Israeli Case. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum Associates

Benítez, I., & Padilla, J. L. (2013). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*. Advance online publication. doi:10.1177/1558689813488245.

Bialostok, S. (2002). Metaphors for literacy: A cultural model of white, middle-class parents. *Linguistics and Education*, *13*(3), 347-371.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313-333.

Bond, L., Moss, P., & Carr, P. (1996). Fairness in large-scale performance assessment. In G.W. Phillips & A. Goldstein (Eds.), Technical issues in large-scale performance assessment (pp. 117–140). Washington, DC: National Center for Education Statistics.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice, 9*(3), 387-99.

Bowles, M., & Stansfield, C. W. (2008). A practical guide to standards-based assessment in the native language. *NLA—LEP Partnership*.

Breithaupt, K., & Zumbo, B. D. (2002). Sample invariance of the structural equation model and the item response model: a case study. *Structural Equation Modeling*, *9*(3), 390-412.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guildford Press.

Bussière, P., Rogers, W. T., Knighton, T., & Cartwright, F. (2004). *The Performance of Canada's Youth in Mathematics, Reading, Science, and Problem Solving: 2003 First Findings for Canadians Aged 15, Highlights*. Statistics Canada.

Campbell, L. (2003). The history of linguistics. In M. Aronoff & J. Rees-Miller (Eds.), *The Handbook of Linguistics* (pp. 81-104). Oxford, UK: Blackwell Publishers.

Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. *Translation today: trends and perspectives*, 205-224.

Cohen, A. (2007). The coming of age for research on test-taking strategies. In J. Fox, M. Wesche, D. Bayliss, . Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 89–111). Ottawa, ON: University of Ottawa Press.

Cohen, Y., Gafni, N., & Hanani, P. (2007). Translating and Adapting a Test, Yet Another Source of Variance; The Standard Error of Translation. Paper presented to the annual meeting of the IAEA, Baku, Azerbaijan. http://www.iaea.info/documents/paper_1162d22ec7.pdf

Cook, L. L., Schmitt-Cascallar, A. P., & Brown, C. (2005). Adapting achievement and aptitude tests: A review of methodological issues. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (p. 171-192*)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Crundwell, R. M. (2005). Alternative Strategies for Large Scale Student Assessment in Canada: Is Value-Added Assessment One Possible Answer. *Canadian Journal of Educational Administration and Policy*, *41*, 1-21.

Cummins, J. (2000). Language, power and pedagogy: *Bilingual children in the cross-fire.* Clevedon, UK: Multilingual Matters Ltd.

Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics, 29*, 90-100.

CTB/McGraw-Hill. (1991). *PARDUX* [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Derrida, (1998). *Of grammatology*. John Hopkins University Press. Baltimore, Maryland.

De Saussure, F. (2011). *Course in general linguistics*. Columbia University Press.

Elosua, P., & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing*, *7*(1), 39-52.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543-553.

Ercikan, K. (2003). Are the English and French versions of the Third International Mathematics and Science Study administered in Canada comparable? Effects of adaptations. *International Journal of Educational Policy, Research and Practice*, *4*, 55–76.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24-35.

Ercikan, K., Domene, J. F., Law, D., Arim, R., Gagnon, F., Lacroix, S. (2004a). *Identifying Sources of DIF Using Think-Aloud Protocols: Comparing Thought Processes of Examinees Taking Tests in English versus in French.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Ercikan, K. Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004b). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301-321.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology. APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 545-569).

Ercikan, K. & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23-35.

Ercikan, K. & McCreith, T. (2002). Disentangling sources of differential item functioning in multi-language assessements. *International Journal of Testing*, *2*,199–215.

Ercikan, K., Oliveri, M. E., & Sandilands, D. (2012). Large-scale assessments of achievement in Canada. *International Guide to Student Achievement*, 456-459.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137-154.

Ercikan, K., Simon, M. & Oliveri, M.E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau, (Eds.). *Improving Large Scale Education Assessment* (pp. 110-121). Routledge.

Ercikan, K., Roth, Simon, Sandilands and Lyons-Thomas. (in press). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied measurement in education*. Manuscript in press.

Fairbairn, S. B., & Fox, J. (2009). Inclusive achievement testing for linguistically and culturally diverse test takers: Essential considerations for test developers and decision makers. *Educational Measurement: Issues and Practice*, *28*(1), 10-24.

Farmer, D. (2008). "My mother is from Russia, My father is from Rwanda" The relationship between immigrant families and the school system in Francophone minority communities. In C. Belkhodja (Ed.), *Special Issue on Immigration and francophone minorities* (pp. 124-127). Retrieved from http://canada.metropolis.net/pdfs/mother_russia_father_rwanda_e.pdf

Foy, P., Brossman, B., & Galia, J. (2013). *Scaling the TIMSS and PIRLS 2011 Achievement Data.* Retrieved from TIMSS & PIRLS International Study Center website: http://timss.bc.edu/pirls2011/international-database.html

Foy, P. & Drucker, T. (2013). *PIRLS 2011 user guide for the international database: PIRLS released passages and items.* Retrieved from TIMSS & PIRLS International Study Center website: http://timss.bc.edu/pirls2011/international- database.html.

Gee, J. P. (2001). Reading as situated language: A sociocognitive perspective. *Journal of Adolescent & Adult Literacy*, *44*(8), 714-725.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological assessment*, *6*(4), 304.

Gierl, M.J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education/Revue canadienne de l'èducation*, *25* (4), 280-96.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*(1), 3–14.

Gierl, M.J. & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement,* Summer 2001, *38* (2),164-187.

Gierl, M.J., Rogers, W.T. & Klinger, D. (1999). *Using statistical and judgmental reviews to identify and interpret translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, CA.

Goldstein, H., & Thomas, S. M. (2008). Reflections on the international comparative surveys debate. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 215-222.

Greenfield, P.M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist,* October 1997, *52*(10), 1115-1124.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, *20*(2), 225-240.

Grisay, A. & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*, 69-86.

Haladyna, T. M., & Downing, S. M. (2004). Construct irrelevant variance in high stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27.

Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguistics and Education*, *5*(2), 93-116.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38*).* Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44(*11), 182-188.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newberry Park, CA: Sage Publications, Inc.

Hammerberg, D. D. (2004). Comprehension instruction for socioculturally diverse classrooms: A review of what we know. *The Reading Teacher*, *57*(7), 648-658.

Hawkins, M.R. (2004). Researching English language and literacy development in schools. *Educational Researcher,* 33(3), 14-25.

Holland, P.W. and Thayer, D.T. 1988: Differential item functioning and the Mantel–Haenszel procedure. In Wainer, H. and Braun, H. I., editors, *Test validity* (pp. 129-45). Hillsdale, NJ: Lawrence Erlbaum.

Hymes, D. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, *23*(2), 8-28.

International Association for the Evaluation of Educational Achievement, (2012). *International database analyzer* (version 3.0). Hamburg, Germany: IEA Data Processing and Research Center.

International Language Testing Association (ILTA) (2000). International Language Testing Association Code of Ethics. Retrieved from: (http://www.iltaonline.com).

International Test Commission (ITC) (2001). *International test commission guidelines for test adaptation.* London. Retrieved from http://www.intestcom.org/itc_projects.

Jang, K.Y. (2010). *Measurement equivalence of Korean and English versions of PISA mathematics item.* Unpublished report, University of British Columbia, Canada.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409-426.

Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation modeling. *Educational Research and Evaluation*, *19*(2-3), 204-222.

Klinger, D. A., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy*, 76. Retrieved from

http://www.umanitoba.ca/publications/cjeap/articles/klinger.html.

Koh, K., & Zumbo, B. D. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods*, *7*(2), 471-477.

Knighton, T., Brochu, P., & Gluszynski, T. (2010). Measuring up: Canadian results of the OECD PISA study: the performance of Canada's youth in reading, mathematics and science 2009: first results for Canadians aged 15.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement, 65*, 935–953.

Labrecque, Chuy, Brochu & Houme, (2012). PIRLS 2011 Canada in Context. Retrieved from Council of Ministers of Education Canada website: http://www.cmec.ca/docs/pirls/PIRLS_2011_Highlights_EN.pdf.

Lazaraton, A., & Taylor, L. (2007). Qualitative research methods in language test development and validation. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 113–137). Ottawa, ON: University of Ottawa Press.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

Maldonado, C. Y., & Geisinger, K. F. (2005). Conversion of the Wechsler Adult Intelligence Scale into Spanish: An early test adaptation effort of considerable consequence. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and*

psychological tests for cross-cultural assessment (pp.213-234). Mahwah, NJ: Lawrence Erlbaum Associates.

Madaus, G., & Clarke, M. (2001). The impact of high-stakes testing on minority students. In M. Kornhaber & G. Orfield (Eds.), *Raising standards or raising barriers: Inequality and high stakes testing in public education* (pp. 85–106). New York: Century Foundation.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*(3), 369-403.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Mesthrie, R. (2008). Sociolinguistics and sociology of language. In B. Spolsky & F.M. Hult (Eds.), *The handbook of educational linguistics* (pp. 66-82). Oxford: Blackwell Publishing.

Mullis, I., Martin, M.O., Kennedy, A.M., Trong, K.L. & Sainsbury, M. (2009). *PIRLS 2011 Assessment Framework.* TIMMS & PIRLS International Study Center, Lynch School of Education, Boston College.

Murat, F., & Rocher, T. (2004). On the methods used for international assessments of educational competences. In JH Moskowitz & M. Stephens (éd.), *Comparing Learning Outcomes: International assessment and education policy (*pp. 190-214*)*. London: Routledge Falmer.

Murata, K. (2007). Unanswered questions: cultural assumptions in text interpretation. *International Journal of Applied Linguistics*, *17*(1), 38-59.

Muthén, L.K. and Muthén, B.O. (1998). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Ochs, E. & Schieffelin, B. (1995). The impact of language socialization on grammatical development. In P. Fletcher & B. MacWhinney, *Handbook of Child Language* (pp. 73-94). Oxford: Blackwell.

Ockey, G. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly, 4*(2), 149–164.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Applied Measurement in Education, 24*, 1–18.

Oliveri, M. E., Olson, B., Ercikan, K., & Zumbo, B.D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, *12*(3), 203-223.

Padilla, A. M. & Medina, A.(1996). Crosscultural sensitivity in assessment: using tests in culturally appropriate ways. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.),*Handbook of multicultural assessment* (pp. 3-28). San Francisco: Jossey-Bass Publishers.

Puhan, G. & Gierl, M.J. (2006). Evaluating the effectiveness of two-stage testing on English and French versions of a science achievement test. *Journal of Cross-Cultural Psychology*, *37*: 136.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517-529.

Rogers, W. T., Lin, J., & Rinaldi, C. M. (2010). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Applied Measurement in Education*, *24*(1), 39-70.

Rorty, R. (1977). Derrida on Language, Being and Abnormal Philosophy. *The Journal of Philosophy, 74* (11), Seventy-Fourth Annual Meeting American Philosophical Association, Eastern Division (Nov., 1977), p. 673-681.

Roth, W.M. (2009). Realizing Vygotsky's program concerning language and thought: tracking knowing (ideas, conceptions, beliefs) in real time. *Language and Education, 23*, (4), 295-311.

Roth, W.M., Oliveri, M.E., Sandilands, D.D., Lyons-Thomas, J. & Ercikan, K. (2013). Investigating linguistic sources of differential item functioning using expert think-aloud protocols in science achievement tests. *International Journal of Science Education, 35*(4), 546-576.

Rueda, R. (2010). 5 Cultural Perspectives in Reading. *Handbook of Reading Research*, *4,* 84.

Rumelhart, D. E. (1994). Toward an interactive model of reading. In R.B. Ruddell, M.R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed., pp.864–894). Newark, DE: International Reading Association.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and social psychology bulletin*, *28*(12), 1629-1646.

Seller, Gafni & Hanani, 2005. Constructing, Adapting, and Validating Admissions Tests in Multiple Languages: The Israeli Case. In R. K. Hambleton, P. F. Merenda & C. D.

Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 297-319). Mahwah, NJ: Lawrence Erlbaum Associates.

Shohamy, E. (2007). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 141–152). Ottawa, ON: University of Ottawa Press.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. *Adapting educational and psychological tests for cross-cultural Assessment* (pp.117-138). Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S. G. (2008). Validity issues in accommodating reading tests. *Jurnal Pendidik dan Pendidikan*, *23*, 81-110.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148–166.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. H. Hambleton, P. F. Merenda, and C. D Spielberger (Eds.) *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 93-116). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations a methodological analysis of translation quality. *Journal of Cross-Cultural Psychology*, *37*(5), 557-567.

Slocum-Gori, S.L. & Zumbo, B.D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102,* 443-461.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record, 108*, 2354–2379.

Solano-Flores, G. (2010). Assessing the cultural validity of assessment practices: An introduction. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 3–21). New York, NY: Routledge.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing*, *9*(2), 78-91.

Solano-Flores, G., Contreras-Niño, L., & Backhoff,, E. (2006). Translation and adaptation of tests: lessons learned and recommendations for countries participating in TIMSS, PISA and other international comparisons. *Revista Electrónica de Investigación Educativa*, *8*(2), 2.

Solano-Flores, G., Contreras-Niño, L. Á., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. In *Research on PISA* (pp. 71-85). Springer Netherlands.

Solano-Flores, G., & Nelson-Barber, S. (2000, April). *Cultural validity of assessments and assessment development procedures.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practices paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing,* 2, 107–129.Steiner, J. & Mahin (1996). Sociocultural approaches to learning and development: A Vygotsky Framework. *Educational Psychologist, 31*(3/4), 191-206.

Steiner, V., & Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educational psychologist*, *31*(3-4), 191-206.

Stobart, G. (2003). The impact of assessment: Intended and unintended consequences. *Assessment in Education, 16*(2), 139–140.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. H. Hambleton, P. F. Merenda, and C. D Spielberger (Eds.) *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 235-263). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ungerleider, C. (2006). Reflections on the use of large-scale student assessment for improving student success. *Canadian Journal of Education*, *29*(3), 873.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*, 1. SAGE Publications, Incorporated.

Van de Vijver F & Tanzer NK. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology 54*, 119–135.

Van de Vijver, F. J., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. H. Hambleton, P. F. Merenda, and C. D Spielberger (Eds.) *Adapting*

Educational and Psychological Tests for Cross-Cultural Assessment (pp. 39-63). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Vera, G. G. (2011). *Languages as factors of reading achievement in PIRLS assessments* (Doctoral dissertation, Université de Bourgogne).

Volante, L. & Jaafar, S.B. (2008). Educational assessment in Canada. *Assessment in education: Principles, policy & practice, 15*:2, 201-210.

Wu, A., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6,* 287-300.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. *Educational Measurement*, *4*, 111-153.

Yildirim, H.H. & Berberoglu, G. (2009). Judgmental and statistical DIF analysis of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*:2, 108-121.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin*, *131*(1), 3.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from

http://educ.ubc.ca/faculty/zumbo/DIF/index.html

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*, 136–147.

Zumbo, B. D. (2005). Structural equation modeling and test validation. *Encyclopedia of statistics in behavioral science*.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223–233.

Zumbo, B. D. (2008). *Statistical methods for investigating item bias in self-report measures.* Universita` degli Studi di Firenze E-prints Archive, Florence, Italy. Retrieved from http://eprints.unifi.it/archive/00001639

Zumbo, B. D., & Koh, K. H. (2005). Manifestation of differences in item-level characteristics in scale-level measurement invariance tests of multi-group confirmatory factor analyses. *Journal of Modern Applied Statistical Methods*, *4*(1), 275-282.

APPENDICES:

Appendix A: English and French Language Passages

Test items were obtained from IEA released passages and items. Source: PIRLS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement, IEA Secretariat, Amsterdam, the Netherlands.

Planning Your Day Hike

- Pick somewhere to go that will be fun and interesting. If in a group, consider everyone when choosing where to go.
- Find out the distance of the hike and how much time it is supposed to take.
- Check out the weather conditions and forecast. Plan and dress the right way for the weather.
- Pack light. Don't make the weight of what you will carry too heavy (see checklist).

Packing Checklist

- ☐ Plenty of water – to keep from getting thirsty
- ☐ Food – high energy snacks or take a picnic lunch
- ☐ First Aid Kit – in case of blisters, scrapes and scratches
- ☐ Insect repellent – to protect from bites (for example – ticks, bees, mosquitoes, and flies).
- ☐ Extra socks – feet may get wet
- ☐ Whistle – important if going alone, three short whistles mean you are in trouble and need assistance
- ☐ Map and compass – very important for more difficult hikes

Keeping Safe on Your Day Hike

- **Start early.** This will give you plenty of time to enjoy your hike and still get back before dark.
- **Stay on hiking trails** unless you know the area.
- **Pace yourself.** Do not hike too quickly so that you can save your energy. When in a group, go only as fast as the slowest member.
- **Be careful where you are walking.** Watch out for things you might trip over like loose rocks, piles of leaves, and sticks. Take care through slippery areas . If you need to go into water, make sure you know how deep it is.
- **Look out for wildlife.** Be careful where you put your feet, when you pick up sticks or rocks and before you sit down. Never approach animals in the wild. They may look cute and harmless, but they can be unpredictable and very protective of their territory.

IMPORTANT: Tell someone about where you are going hiking and when you expect to return. This could help in case something happens and you get into trouble. Let him or her know when you get back.

Most of all, don't forget to have fun on your hike. Enjoy being outdoors. Look at all the interesting things around you. Learn to identify new places, plants, and animals. Appreciate the beauty of the land and nature, and get good healthy exercise too!

Discover the Fun of Day Hiking

Looking for something fun and interesting to do at home or on holiday?

One of the greatest ways to enjoy the outdoors is hiking, and day hiking is the most popular kind. It doesn't have to take much time or require any special equipment.

Day Hiking Is Fun and Good Exercise!

You are in charge! You can choose where you want to go, how long you want to be gone, and how fast you want to go. You can simply stroll along enjoying nature or challenge yourself with difficult and steep hiking trails. It is up to you!

See interesting new things! Hiking can take you places that cannot be seen any other way. You can go to beautiful areas and see spectacular views. Or you can go to remote areas that may have hidden valleys, waterfalls, or caves. Hiking can give you a chance to see plants, birds, and animals that live in the wild. You might even see remains of buildings and things that belonged to people who lived long ago.

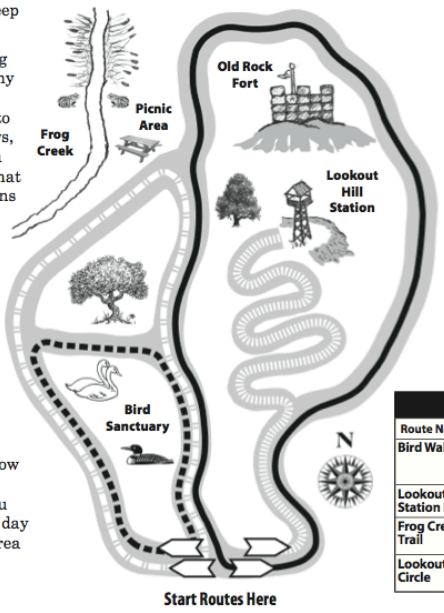
Keep physically fit! Walking is an excellent way to exercise, so hiking on a regular basis will help to keep you healthy. It provides time to think and can be relaxing. Hiking is a great way to spend time with your friends and family or to just spend a little time by yourself studying and enjoying nature.

Explore Lookout Hill

The map and map key for Lookout Hill show how you can choose the day hike that you would like best and the kinds of things you can see and do. It gives you an idea about day hiking in case you want to find a hiking area near where you live.

Lookout Hill

A Hike Full of Adventures



Choose which route to take!

Use one of our suggestions, or make up a route of your own.

Map Key

Route Name	Route	Time	Level	Description
Bird Walk		2 hours	Easy, Wheelchair access	Loop around the bird sanctuary
Lookout Station Hike		2 to 2.5 hours each way	Hard	Climb Lookout Hill and see the view
Frog Creek Trail		3 hours	Medium	Hike to Frog Creek Picnic Area
Lookout Hill Circle		5 hours	Medium	Hike around Lookout Hill to the Old Rock Fort

Questions Discover the Fun of Day Hiking

Take out the leaflet called Discover the Fun of Day Hiking.
The questions in this section are about this leaflet.

Raise your hand if you do not have the leaflet.

1. What is the **main** message the leaflet gave you about hiking?

   Ⓐ It is expensive and dangerous.

   Ⓑ It is the best way to see animals.

   Ⓒ It is healthy and fun.

   Ⓓ It is only for experts.

2. Give **two** interesting things the leaflet said you might see on a day hike.

   1. _____

   2. _____

3. What are two things the leaflet told you to keep in mind when you are hiking in a **group**?

   1. _____

   _____

   2. _____

   _____

4. Which section of the leaflet told you to wear the right clothes for the weather?

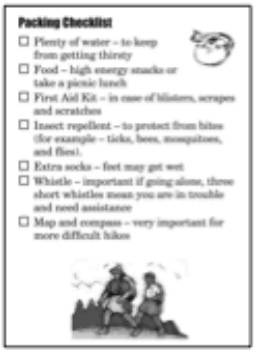
(A) Discover the Fun of Day Hiking

(B) Planning Your Day Hike

(C) Packing Checklist

(D) Keeping Safe on Your Day Hike

Look at the section called *Packing Checklist*. Use it to answer Questions 5 and 6.


Packing Checklist

- ☐ Plenty of water – to keep from getting thirsty
- ☐ Food – high energy snacks or take a picnic lunch
- ☐ First Aid Kit – in case of blisters, scrapes and scratches
- ☐ Insect repellent – to protect from bites (for example – ticks, bees, mosquitoes, and flies).
- ☐ Extra socks – feet may get wet
- ☐ Whistle – important if going alone, three short whistles mean you are in trouble and need assistance
- ☐ Map and compass – very important for more difficult hikes

5. Why should you take extra socks on your hike?

(A) feet may get wet

(B) weather may get cold

(C) in case of blisters

(D) for a friend

6. What should you do if you get in trouble while on your hike?

(A) have a high energy snack

(B) blow your whistle three times

(C) put on more insect repellent

(D) yell for help as loud as you can

Look at the section called *Keeping Safe*. Use it to answer Questions 7 and 8.

Keeping Safe on Your Day Hike

- **Start early**. This will give you plenty of time to enjoy your hike and still get back before dark.
- **Stay on hiking trails** unless you know the area.
- **Pace yourself**. Do not hike too quickly so that you can save your energy. When in a group, go only as fast as the slowest member.
- **Be careful where you are walking**. Watch out for things you might trip over like loose rocks, piles of leaves, and sticks. Take care through slippery areas . If you need to go into water, make sure you know how deep it is.
- **Look out for wildlife.** Be careful where you put your feet, when you pick up sticks or rocks, and before you sit down. Never approach animals in the wild. They may look cute and harmless, but they can be unpredictable and very protective of their territory.

IMPORTANT: Tell someone about where you are going hiking and when you expect to return. This could help in case something happens and you get into trouble. Let him or her know when you get back.

7. What should you do to avoid getting tired too soon?

   (A) start early

   (B) stay on hiking trails

   (C) pace yourself

   (D) be careful where you walk

8. Why is it important to tell someone when you plan to return from your hike?

   1 _____

   _____

Use the information about the *Lookout Hill Hike* to answer Questions 9 through 12.

9. Which route would you choose if you wanted to take the shortest hike?

   (A) Bird Walk

   (B) Lookout Station Hike

   (C) Frog Creek Trail

   (D) Lookout Hill Circle

10. Which kind of people would be most able to go on the Lookout Station Hike?

   (A) people who are in a hurry

   (B) people who have small children

   (C) people who like to watch birds

   (D) people who are fit and strong

11. What are **two** things you can learn by studying the map key?

1 1. _____

_____

1 2. _____

_____

12. Use the map of Lookout Hill and the map key to plan a hike. Check which route you would choose.

_____ Bird Walk

_____ Lookout Station Hike

_____ Frog Creek Trail

_____ Lookout Hill Circle

Give **two** reasons from the leaflet why you chose this route.

1 1. _____

_____

1 2. _____

Planification de ta randonnée

- Choisis un endroit qui sera plaisant et intéressant pour tout le monde. Tiens compte de chaque personne dans ton groupe pour choisir où aller.
- Vérifie la longueur de la randonnée et combien de temps elle est censée prendre.
- Vérifie les conditions et les prévisions météorologiques. Prévois un habillement approprié.
- Ne te charge pas trop. Choisis du matériel qui ne soit pas trop lourd (voir la liste de vérification).

Liste de vérification

- ☐ Suffisamment d'eau – pour éviter de devenir assoiffé.
- ☐ Nourriture – des collations hautement énergétiques ou un pique-nique.
- ☐ Trousse de premiers soins – en cas d'ampoule, d'égratignure ou de petite blessure.
- ☐ Chasse-insectes – pour se protéger des piqûres (par exemple, des tiques, des abeilles, des moustiques et des mouches).
- ☐ Bas de rechange – les pieds peuvent se mouiller.
- ☐ Sifflet – important si tu pars sans compagnie; trois brefs coups de sifflet signifient que tu as des problèmes et que tu as besoin d'aide.
- ☐ Carte et boussole – très important pour les randonnées plus difficiles.

Être en sécurité lors de ta randonnée

- **Pars tôt.** Cela te donnera assez de temps pour apprécier la randonnée et revenir avant la noirceur.
- **Reste sur les sentiers identifiés,** sauf si tu connais la région.
- **Suis ton rythme.** Ne marche pas trop vite afin d'économiser ton énergie. Ne va pas plus vite que la personne la plus lente du groupe.
- **Fais attention où tu marches.** Prends soin de ne pas trébucher sur des rochers instables, des tas de feuilles ou des morceaux de bois par exemple. Ne te presse pas dans les endroits glissants. Si tu dois entrer dans l'eau, sois sûr que tu en connais la profondeur.
- **Fais attention à la flore et à la faune.** Regarde où tu mets les pieds; fais attention quand tu ramasses des bâtons ou des cailloux, et avant de t'asseoir. Ne t'approche jamais d'animaux sauvages! Ceux-ci peuvent sembler mignons et sans danger, mais ils peuvent être imprévisibles et se montrer très protecteurs de leur territoire.

IMPORTANT: Dis à quelqu'un où tu veux aller faire ta randonnée et quand tu prévois revenir. Cela pourrait être utile s'il arrivait quelque chose et que tu avais des ennuis. Avertis cette personne lorsque tu seras rentré.

Surtout, n'oublie pas de t'amuser lors de ta randonnée. Profite de la nature. Regarde tout ce qu'il y a d'intéressant autour de toi. Apprends à identifier de nouveaux endroits, de nouvelles plantes et de nouveaux animaux. Apprécie la beauté du lieu et de la nature, tout en prof tant d'une activité physique saine!

Découvre les joies de la randonnée

Est-ce que tu cherches quelque chose d'amusant et d'intéressant à faire à la maison ou en vacances?

Une des meilleures façons de profiter du plein air est la randonnée; les excursions d'un jour sont celles qui remportent le plus de succès. Cela ne prend pas nécessairement beaucoup de temps et ne demande pas d'équipement particulier.

La randonnée est amusante et fait faire de l'exercice!

Tu prends toi-même les choses en main! Tu peux choisir où tu veux aller, combien de temps tu veux marcher et à quel rythme tu veux aller. Tu peux simplement te promener en appréciant la nature ou te donner un défi avec des chemins raides et difficiles. Cela ne dépend que de toi!

Découvre des choses intéressantes! La randonnée peut t'amener à des endroits qui ne peuvent pas être vus autrement. Tu peux visiter des régions magnifiques et voir des paysages spectaculaires. Ou tu peux aller dans des endroits reculés qui possèdent peut-être des vallées, des chutes d'eau ou des grottes cachées. La randonnée te donne la possibilité de voir des plantes, des oiseaux et des animaux sauvages. Peut-être que tu verras des ruines ou des objets ayant appartenu à des peuples qui ont vécu il y a longtemps.

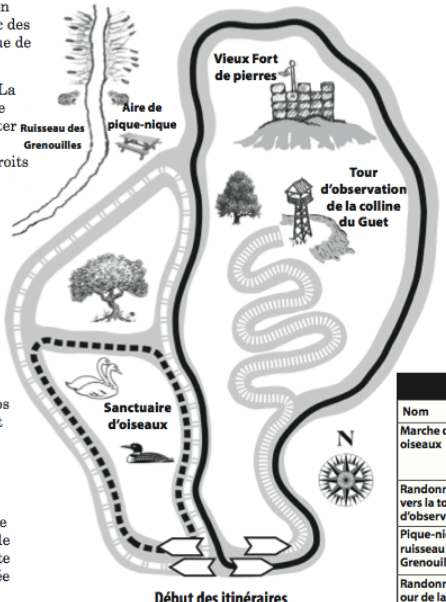
Garde la forme! La marche est un excellent moyen de faire de l'exercice. Donc, faire régulièrement des randonnées t'aidera à te maintenir en bonne santé. Cela donne le temps de penser et peut être relaxant. Grâce à la randonnée, tu peux passer du temps avec tes amis et ta famille ou tout simplement prendre un peu de temps pour toi-même en observant et en appréciant la nature.

Explore la colline du Guet

La Carte de randonnée de la colline du Guet te montre comment choisir la marche qui t'irait le mieux et ce que tu peux y voir et y faire. Elle te donne ainsi des idées pour faire une randonnée dans une région près de chez toi.

La colline du Guet :

une randonnée pleine d'aventures



Choisis ton itinéraire!

Utilise une de nos suggestions ou crée ton propre itinéraire.

Légende				
Nom	**Itinéraire**	**Durée**	**Niveau**	**Description**
Marche des oiseaux	■ ■ ■	2 heures	Facile Accessible en fauteuil roulant	Faire le tour du sanctuaire d'oiseaux
Randonnée vers la tour d'observation	▮▮▮▮▮▮▮	2 à 2,5 heures par trajet	Difficile	Grimper la colline du Guet et admirer la vue depuis la tour d'observation
Pique-nique au ruisseau des Grenouilles	■ ▮ ■	3 heures	Moyen	Faire une randonnée jusqu'à l'aire de pique-nique du ruisseau des Grenouilles
Randonnée autour de la colline du Guet	▬	5 heures	Moyen	Faire une randonnée autour de la colline du Guet jusqu'au Vieux Fort de pierres

Questions Découvre les joies de la randonnée

Prends le dépliant intitulé « Découvre les joies de la randonnée ». Les questions de cette section se rapportent à ce dépliant.

Lève la main si tu n'as pas ce dépliant.

1. Quelle est l'idée **principale** de ce dépliant au sujet du plein air?

   (A) C'est cher et dangereux.

   (B) C'est le meilleur moyen de voir des animaux

   (C) C'est bon pour la santé et c'est amusant.

   (D) C'est réservé aux experts.

2. Nomme **deux** choses intéressantes que tu pourrais voir pendant une randonnée, d'après ce dépliant.

   1. _____

   2. _____

3. Nomme deux points, décrits dans le dépliant, dont tu dois tenir compte lorsque tu pars en randonnée en **groupe**.

   1. _____

   _____

   2. _____

   _____

4. Dans quelle section du dépliant indique-t-on de porter un habillement approprié aux conditions météorologiques?

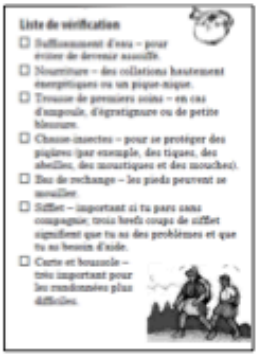
(A) Découvre les joies de la randonnée.

(B) Planification de ta randonnée.

(C) Liste de vérification.

(D) Être en sécurité lors de ta randonnée.

Regarde la section intitulée *Liste de vérification*. Sers-toi de cette section pour répondre aux questions 5 et 6.


Liste de vérification

☐ Suffisamment d'eau – pour éviter de devenir assoiffé.

☐ Nourriture – des collations hautement énergétiques ou un pique-nique.

☐ Trousse de premiers soins – en cas d'ampoule, d'égratignure ou de petite blessure.

☐ Chasse-insectes – pour se protéger des piqûres (par exemple, des tiques, des abeilles, des moustiques et des mouches).

☐ Bas de rechange – les pieds peuvent se mouiller.

☐ Sifflet – important si tu pars sans compagnie; trois brefs coups de sifflet signifient que tu as des problèmes et que tu as besoin d'aide.

☐ Carte et boussole – très important pour les randonnées plus difficiles.

5. Pourquoi devrais-tu apporter des bas de rechange lorsque tu pars en randonnée?

(A) Tu peux te mouiller les pieds.

(B) Il peut faire froid.

(C) En cas d'ampoules.

(D) Pour prêter à un ami.

6. Que dois-tu faire si tu as des problèmes pendant ta randonnée?

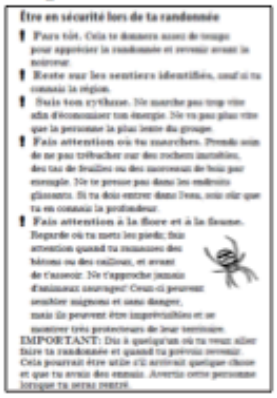
(A) Avoir une collation hautement énergétique.

(B) Siffler trois coups avec ton sifflet.

(C) Appliquer plus de chasse-insectes.

(D) Crier le plus fort possible pour appeler à l'aide.

Regarde la section intitulée *Être en sécurité lors de ta randonnée*. Sers-toi de cette section pour répondre aux questions 7 et 8.



Être en sécurité lors de ta randonnée

- Pars tôt. Cela te donnera assez de temps pour apprécier la randonnée et revenir avant la noirceur.
- Reste sur les sentiers identifiés, sauf si tu connais la région.
- Suis ton rythme. Ne marche pas trop vite afin d'économiser ton énergie. Ne va pas plus vite que la personne la plus lente du groupe.
- Fais attention où tu marches. Prends soin de ne pas trébucher sur des rochers instables, des tas de feuilles ou des morceaux de bois par exemple. Ne te presse pas dans les endroits glissants. Si tu dois entrer dans l'eau, sois sûr que tu en connais la profondeur.
- Fais attention à la flore et à la faune. Regarde où tu mets les pieds; fais attention quand tu ramasses des bâtons ou des cailloux, et avant de t'asseoir. Ne t'approche jamais d'animaux sauvages! Ceux-ci peuvent semblent mignons et sans danger, mais ils peuvent être imprévisibles et se montrer très protecteurs de leur territoire.

IMPORTANT: Dis à quelqu'un où tu veux aller faire ta randonnée et quand tu prévois revenir. Cela pourrait être utile s'il arrivait quelque chose et que tu avais des ennuis. Avertis cette personne lorsque tu seras rentré.

7. Que dois-tu faire pour éviter de te fatiguer trop rapidement?

   (A) Partir tôt.

   (B) Rester sur les sentiers identifiés.

   (C) Suivre ton rythme.

   (D) Faire attention où tu marches.

8. Pourquoi est-il important de dire à quelqu'un quand tu prévois revenir?

_____

_____

Sers-toi de l'information sur la *Randonnée de la colline du Guet* pour répondre aux questions 9 à 12.

9. Quel itinéraire choisirais-tu si tu voulais faire une randonnée qui soit la plus courte possible?

   (A) Marche des oiseaux.

   (B) Randonnée vers la tour d'observation.

   (C) Pique-nique au ruisseau des Grenouilles.

   (D) Randonnée autour de la colline du Guet.

10. Quelles personnes seraient les plus habiles pour faire la Randonnée vers la tour d'observation?

   (A) Des gens pressés.

   (B) Des gens accompagnés de petits enfants.

   (C) Des gens qui aiment observer les oiseaux.

   (D) Des gens en forme et forts.

11. Nomme **deux** choses que tu peux apprendre en observant le tableau qui accompagne la carte.

1. _____

_____

2. _____

_____

12. Sers-toi de la carte et du tableau sur la colline du Guet pour planifier une randonnée. Coche l'itinéraire de ton choix.

_____ Marche des oiseaux.

_____ Randonnée vers la tour d'observation.

_____ Pique-nique au ruisseau des Grenouilles.

_____ Randonnée autour de la colline du Guet.

Donne **deux** raisons, tirées du dépliant, pour lesquelles tu choisis cet itinéraire.

1. _____

_____

2. _____

Enemy Pie

by Derek Munson
illustrated by Tara Calahan King

It was a perfect summer until Jeremy Ross moved in right next door to my best friend Stanley. I did not like Jeremy. He had a party and I wasn't even invited. But my best friend Stanley was.

I never had an enemy until Jeremy moved into the neighbourhood. Dad told me that when he was my age, he had enemies, too. But he knew of a way to get rid of them.

Dad pulled a worn-out scrap of paper from a recipe book.

"Enemy Pie," he said, satisfied.

You may be wondering what exactly is in Enemy Pie. Dad said the recipe was so secret, he couldn't even tell me. I begged him to tell me something—anything.

"I will tell you this, Tom," he said to me. "Enemy Pie is the fastest known way to get rid of enemies."

This got me thinking. What kinds of disgusting things would I put into Enemy Pie? I brought Dad earthworms and rocks, but he gave them right back.

I went outside to play. All the while, I listened to the sounds of my dad in the kitchen. This could be a great summer after all.

I tried to imagine how horrible Enemy Pie must smell. But I smelled something really good. As far as I could tell, it was coming from our kitchen. I was confused.

I went inside to ask Dad what was wrong. Enemy Pie shouldn't smell this good. But Dad was smart. "If it smelled bad, your enemy would never eat it," he said. I could tell he'd made this pie before.

The oven buzzer rang. Dad put on oven mitts and pulled out the pie. It looked good enough to eat! I was beginning to understand.

But still, I wasn't sure how this Enemy Pie worked. What exactly did it do to enemies? Maybe it made their hair fall out, or their breath stinky. I asked Dad, but he was no help.

While the pie cooled, Dad filled me in on my job.

He whispered. "In order for it to work, you need to spend a day with your enemy. Even worse, you have to be nice to him. It's not easy. But that's the only way that Enemy Pie can work. Are you sure you want to do this?"

Of course I was.

All I had to do was spend one day with Jeremy, then he'd be out of my life. I rode my bike to his house and knocked on the door.

When Jeremy opened the door, he seemed surprised.

"Can you come out and play?" I asked.

He looked confused. "I'll go ask my mom," he said. He came back with his shoes in his hand.

We rode bikes for awhile, then ate lunch. After lunch we went over to my house.

It was strange, but I was having fun with my enemy. I couldn't tell Dad that, since he had worked so hard to make the pie.

We played games until my dad called us for dinner.

Dad had made my favourite food. It was Jeremy's favourite, too! Maybe Jeremy wasn't so bad after all. I was beginning to think that maybe we should forget about Enemy Pie.

"Dad", I said, "It sure is nice having a new friend." I was trying to tell him that Jeremy was no longer my enemy. But Dad only smiled and nodded. I think he thought I was just pretending.

But after dinner, Dad brought out the pie. He dished up three plates and passed one to me and one to Jeremy.

"Wow!" Jeremy said, looking at the pie.

I panicked. I didn't want Jeremy to eat Enemy Pie! He was my friend!

"Don't eat it!" I cried. "It's bad!"

Jeremy's fork stopped before reaching his mouth. He looked at me funny. I felt relieved. I had saved his life.

"If it's so bad," Jeremy asked, "then why has your dad already eaten half of it?"

Sure enough, Dad was eating Enemy Pie.

"Good stuff," Dad mumbled. I sat there watching them eat. Neither one of them was losing any hair! It seemed safe, so I took a tiny taste. It was delicious!

After dessert, Jeremy invited me to come over to his house the next morning.

As for Enemy Pie, I still don't know how to make it. I still wonder if enemies really do hate it or if their hair falls out or their breath turns bad. But I don't know if I'll ever get an answer, because I just lost my best enemy.

Questions Enemy Pie

1. Who is telling the story?

   (A) Jeremy

   (B) Dad

   (C) Stanley

   (D) Tom

2. At the beginning of the story, why did Tom think Jeremy was his enemy?

   _____

   _____

3. Write **one** ingredient that Tom thought would be in Enemy Pie.

   _____

4. Find the part of the story next to the picture of a piece of pie:  .
Why did Tom think it could be a great summer after all?

(A) He liked playing outside.

(B) He was excited about Dad's plan.

(C) He made a new friend.

(D) He wanted to taste Enemy Pie.

5. How did Tom feel when he first smelled Enemy Pie? Explain why he felt this way.

_____

_____

_____

2 1 0 | 8 9

6. What did Tom think could happen when his enemy ate Enemy Pie?
Write **one** thing.

_____

_____

1 0 | 8 9

7. What were the **two** things Tom's dad told Tom to do for Enemy Pie to work?

2

_____

_____

8. Why did Tom go to Jeremy's house?

(A) To invite Jeremy to dinner.

(B) To ask Jeremy to leave Stanley alone.

(C) To invite Jeremy to play.

(D) To ask Jeremy to be his friend.

9. What surprised Tom about the day he spent with Jeremy?

1

_____

_____

10. At dinner, why did Tom begin to think he and his dad should forget about Enemy Pie?

(A) Tom did not want to share dessert with Jeremy.

(B) Tom did not think Enemy Pie would work.

(C) Tom was beginning to like Jeremy.

(D) Tom wanted to keep Enemy Pie a secret.

11. How was Tom feeling when Dad passed the piece of Enemy Pie to Jeremy?

(A) alarmed

(B) satisfied

(C) surprised

(D) confused

12. What was it about Enemy Pie that Dad kept secret?

(A) It was a normal pie.

(B) It tasted disgusting.

(C) It was his favourite food.

(D) It was a poisonous pie.

13. Look at this sentence from the end of the story:

"After dessert, Jeremy invited me to come over to his house the next morning."

What does this suggest about the boys?

(A) They are still enemies.

(B) They do not like to play at Tom's house.

(C) They wanted to eat some more Enemy Pie.

(D) They might be friends in the future.

14. Use what you have read to explain why Tom's dad really made Enemy Pie.

15. What kind of person is Tom's dad? Give an example of what he did in the story that shows this.

2

16. What lesson might you learn from this story?

1

La tarte des ennemis

Derek Munson
Illustrations de Tara Calahan King

L'été avait été parfait jusqu'à ce que Jérémie Leroux emménage juste à côté de chez Sammy, mon meilleur ami. Je n'aimais pas Jérémie. Un jour, il avait organisé une fête et ne m'avait même pas invité. Mais il avait invité Sammy, mon meilleur ami.

Je n'avais jamais eu d'ennemi avant que Jérémie arrive dans le quartier. Papa m'a dit que lui aussi avait eu des ennemis quand il avait mon âge. Et qu'il connaissait un moyen de s'en débarrasser.

D'un livre de recettes, il a tiré un bout de papier très abîmé.

« La tarte des ennemis », a-t-il annoncé, l'air satisfait.

Vous vous demanderez sans doute ce qu'il y a dans une tarte des ennemis. En fait, papa m'a dit que la recette était tellement secrète qu'il ne pouvait même pas me le dire. Je l'ai supplié de m'en dire quelque chose… n'importe quoi!

« D'accord, Thomas. Voici ce que je peux t'en dire : la tarte des ennemis est le moyen le plus rapide de se débarrasser de ses ennemis. »

Voilà qui m'a fait réfléchir. Quelles choses dégoûtantes pouvais-je bien mettre dans cette tarte des ennemis? J'ai apporté à papa des vers de terre et des cailloux. Mais il me les a tout de suite rendus.

Je suis sorti jouer, mais j'écoutais papa qui s'agitait dans la cuisine. L'été serait peut-être agréable, après tout.

J'ai essayé d'imaginer l'odeur horrible de la tarte des ennemis. À ce moment, pourtant, je sentais une bonne odeur. Et d'après mon nez, cette odeur venait de notre cuisine. Je ne savais plus que penser.

Je suis rentré pour demander à papa ce qui n'allait pas. La tarte des ennemis ne devrait pas sentir aussi bon. Mais papa était plutôt futé. « Si la tarte sentait mauvais, ton ennemi n'en mangerait pas. » J'ai compris qu'il avait déjà cuisiné cette tarte auparavant.

La sonnerie du fourneau a retenti. Papa a enfilé les gants de cuisine et a sorti la tarte. Elle avait l'air délicieuse! Je commençais à comprendre.

Je n'étais toujours pas certain de ses effets. Qu'est-ce qu'elle faisait exactement aux ennemis? Elle leur faisait perdre leurs cheveux? Elle leur donnait mauvaise haleine? J'ai demandé à papa, mais sans résultat.

Tandis que la tarte refroidissait, papa m'a expliqué la suite du plan.

Il parlait à voix basse. « Pour que ça fonctionne, tu dois passer une journée entière avec ton ennemi. Pire : tu dois être gentil avec lui. Ce ne sera pas facile, mais c'est le seul moyen d'obtenir les résultats attendus. Es-tu certain de vouloir le faire? »

Tu parles que j'étais certain!

Il me suffisait donc de passer une journée avec Jérémie pour qu'il disparaisse ensuite de ma vie. Je me suis rendu chez lui à vélo et j'ai frappé à sa porte.

En m'ouvrant, il a semblé surpris.

« Peux-tu jouer avec moi? », lui ai-je demandé.

Il avait l'air de ne pas comprendre. « Je vais demander à ma mère. », a-t-il répondu. Puis il est revenu avec ses chaussures à la main.

Nous nous sommes baladés à vélo, nous avons pique-niqué puis nous sommes allés chez moi.

C'était étrange mais je m'amusais bien avec mon ennemi. Je ne pouvais pas en parler à papa : il avait travaillé si fort pour cuisiner la tarte!

Nous avons joué jusqu'à ce que papa nous appelle pour le repas.

Papa avait cuisiné mon plat favori, qui était aussi le plat favori de Jérémie! Peut-être que Jérémie n'était pas si mal, après tout. Je commençais à penser qu'il valait peut-être mieux éviter la tarte des ennemis.

« Papa », ai-je dit, « c'est tellement agréable d'avoir un nouvel ami. » En fait, j'essayais de lui faire comprendre que Jérémie n'était plus mon ennemi. Papa s'est contenté de sourire et de hocher la tête. Il a dû croire que je faisais semblant.

À la fin du repas, papa a apporté la tarte. Il a coupé trois parts, en a posé une devant Jérémie et une autre devant moi.

« Super! », dit Jérémie en voyant la tarte.

J'ai paniqué. Je ne voulais pas que Jérémie mange de la tarte des ennemis. C'était mon ami!

« Ne mange pas! » ai-je crié. « C'est mauvais! »

Jérémie a arrêté son geste avant que sa fourchette n'atteigne sa bouche. Il m'a jeté un regard étrange. Quel soulagement! Je lui avais sauvé la vie.

« Si c'est tellement mauvais, pourquoi ton père a-t-il déjà mangé la moitié de sa part? », a demandé Jérémie.

C'était bien vrai : papa mangeait la tarte des ennemis.

« Miam, drôlement bon », a marmonné papa. Je les ai regardés manger. Aucun des deux ne perdait ses cheveux… La tarte semblait sans danger. J'en ai donc pris une toute petite bouchée. C'était délicieux!

Après le dessert, Jérémie m'a invité à passer chez lui le lendemain matin.

Quant à la tarte des ennemis, je ne sais toujours pas comment la cuisiner. Je me demande encore si les ennemis la détestent ou s'ils perdent leurs cheveux ou si leur haleine devient horrible. Je ne sais pas non plus si je vais connaître la réponse un jour, puisque que je viens de perdre mon meilleur ennemi.

Questions La tarte des ennemis

1. Qui raconte l'histoire?

   (A) Jérémie

   (B) Le père

   (C) Sammy

   (D) Thomas

2. Au début de l'histoire, pourquoi Thomas pense-t-il que Jérémie est son ennemi?

   1

   _____

   _____

3. Écris **un** des ingrédients que Thomas s'attendait à trouver dans la tarte des ennemis.

   1

   _____

4. Relis le passage de l'histoire qui se trouve à côté de l'illustration d'une pointe de tarte  Pourquoi Thomas pensait-il que l'été serait peut-être agréable, après tout?

Ⓐ Parce qu'il aimait jouer dehors.

Ⓑ Parce qu'il était content du plan proposé par son père.

Ⓒ Parce qu'il avait un nouvel ami.

Ⓓ Parce qu'il voulait goûter la tarte des ennemis.

5. Comment Thomas a-t-il réagi quand il a senti l'odeur de la tarte des ennemis? Explique pourquoi il a réagi de cette façon.

_____

_____

_____

6. Selon Thomas, qu'est-ce qui pouvait arriver à son ennemi s'il mangeait de la tarte des ennemis?
Écris **une** conséquence.

_____

_____

7. Le père a fait **deux** recommandations à Thomas pour que la tarte des ennemis fonctionne. Lesquelles?

2

_____

_____

8. Pourquoi Thomas est-il allé chez Jérémie?

(A) Pour inviter Jérémie à manger.

(B) Pour demander à Jérémie de laisser Sammy en paix.

(C) Pour inviter Jérémie à jouer.

(D) Pour demander à Jérémie d'être son ami.

9. De quoi Thomas était-il surpris au cours de la journée passée avec Jérémie?

1

_____

_____

12. Quel secret le père avait-il gardé à propos de la tarte des ennemis?

(A) C'était une tarte comme les autres.

(B) Elle avait un goût horrible.

(C) C'était son plat favori.

(D) La tarte était empoisonnée.

13. Relis cette phrase, qui se trouve vers la fin de l'histoire :

> « Après le dessert, Jérémie m'a invité à passer chez lui le lendemain matin. »

Qu'est-ce que cette phrase permet de conclure au sujet des deux garçons?

(A) Qu'ils sont restés ennemis.

(B) Qu'ils n'aiment pas jouer chez Thomas.

(C) Qu'ils voulaient encore manger de la tarte des ennemis.

(D) Qu'ils pourraient devenir amis à l'avenir.

14. Aide-toi du texte pour expliquer pourquoi le papa de Thomas a réellement cuisiné la tarte des ennemis.

1 _____

_____

_____

15. Quel genre de personne est le père de Thomas? Trouve un exemple dans le récit pour expliquer ta réponse.

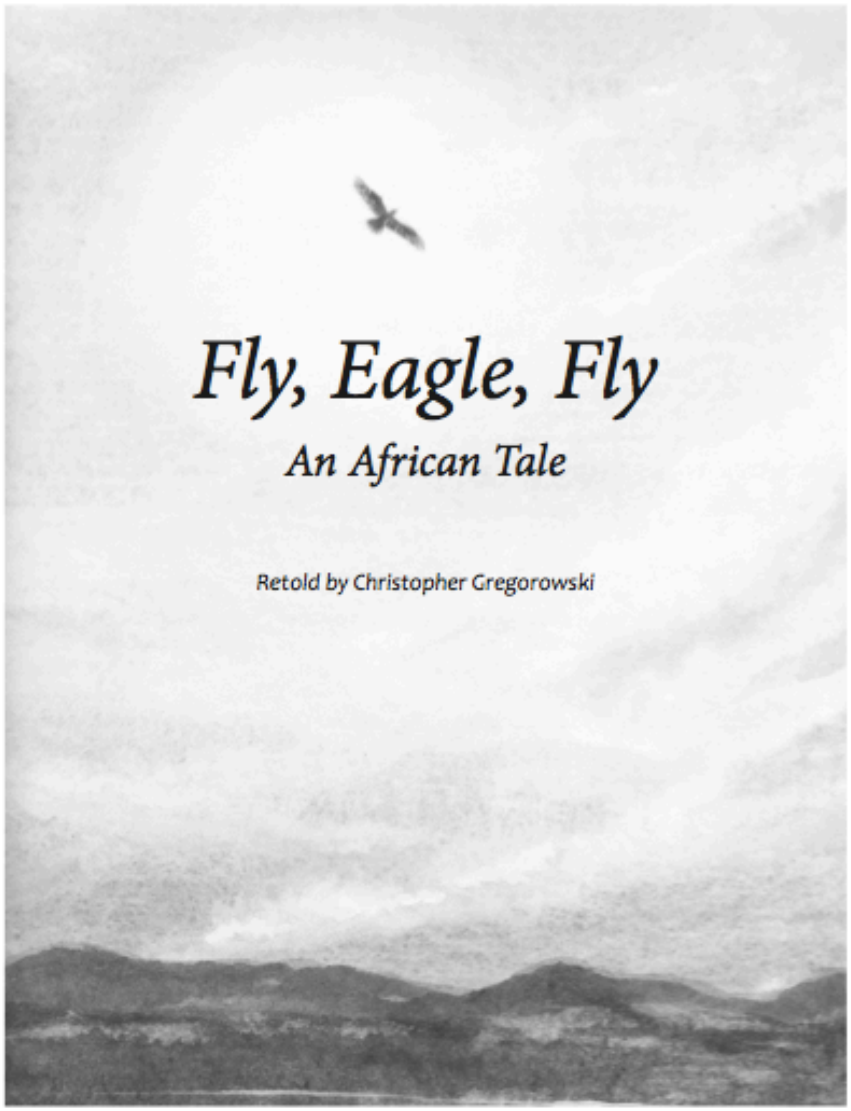
2

2 1 0 8 9

16. Quelle leçon peux-tu tirer de cette histoire?

1

1 0 8 9


Fly, Eagle, Fly
An African Tale
Retold by Christopher Gregorowski

A farmer went out one day to search for a lost calf. The herders had returned without it the evening before. And that night there had been a terrible storm.

He went to the valley and searched by the riverbed, among the reeds, behind the rocks and in the rushing water.

He climbed the slopes of the high mountain with its rocky cliffs. He looked behind a large rock in case the calf had huddled there to escape the storm. And that was where he stopped. There, on a ledge of rock, was a most unusual sight. An eagle chick had hatched from its egg a day or two earlier, and had been blown from its nest by the terrible storm.

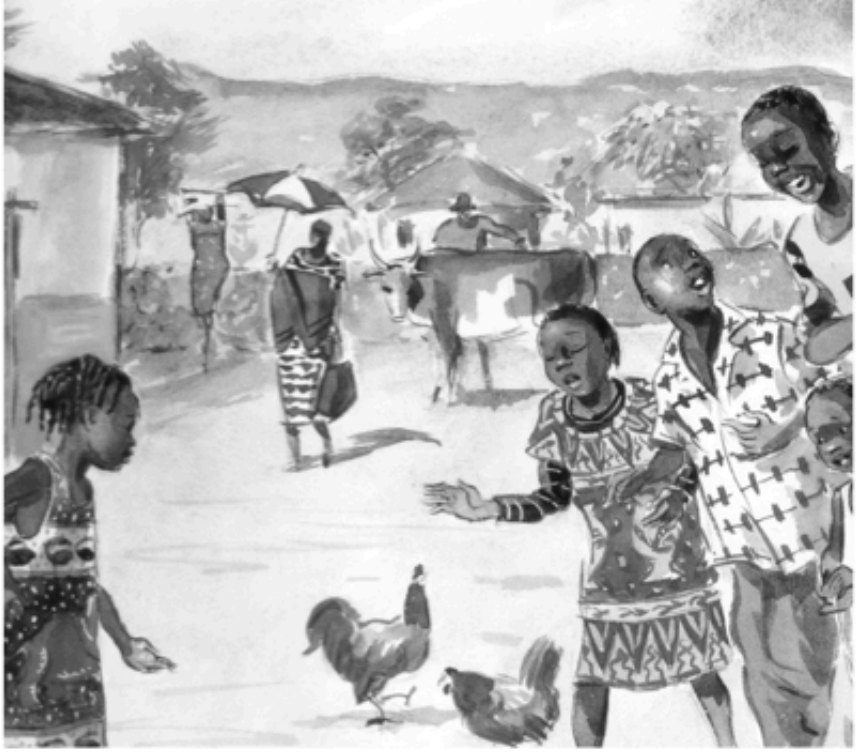
He reached out and cradled the chick in both hands. He would take it home and care for it.

He was almost home when the children ran out to meet him.

"The calf came back by itself!" they shouted.

The farmer was very pleased. He showed the eagle chick to his family, then placed it carefully in the chicken house among the hens and chicks.

"The eagle is the king of the birds," he said, "but we shall train it to be a chicken."



So, the eagle lived among the chickens, learning their ways. As it grew, it began to look quite different from any chicken they had ever seen.

One day a friend dropped in for a visit. The friend saw the bird among the chickens.

"Hey! That is not a chicken. It's an eagle!"

The farmer smiled at him and said, "Of course it's a chicken. Look—it walks like a chicken, it eats like a chicken. It thinks like a chicken. Of course it's a chicken."

But the friend was not convinced. "I will show you that it is an eagle," he said.

The farmer's children helped his friend catch the bird. It was fairly heavy, but the farmer's friend lifted it above his head and said, "You are not a chicken but an eagle. You belong not to the earth but to the sky. Fly, Eagle, fly!"

The bird stretched out its wings, looked about, saw the chickens feeding, and jumped down to scratch with them for food.

"I told you it was a chicken," the farmer said, and he roared with laughter.

Very early the next morning the farmer's dogs began to bark. A voice was calling outside in the darkness. The farmer ran to the door. It was his friend again. "Give me another chance with the bird," he begged.

"Do you know the time? It is long before dawn."

"Come with me. Fetch the bird."

Reluctantly, the farmer picked up the bird, which was fast asleep among the chickens. The two men set off, disappearing into the darkness.

"Where are we going?" asked the farmer sleepily.

"To the mountains where you found the bird."

"And why at this ridiculous time of the night?"

"So that our eagle may see the sun rise over the mountain and follow it into the sky where it belongs."

They went into the valley and crossed the river, the friend leading the way. "Hurry," he said, "for the dawn will arrive before we do."

The first light crept into the sky as they began to climb the mountain. The wispy clouds in the sky were pink at first, and then began to shimmer with a golden brilliance. Sometimes their path was dangerous as it clung to the side of the mountain, crossing narrow shelves of rock and taking them into dark crevices and out again. At last he said, "This will do." He looked down the cliff and saw the ground hundreds of metres below. They were very near the top.

Carefully, the friend carried the bird onto a ledge. He set it down so that it looked toward the east, and began talking to it. The farmer chuckled. "It talks only chicken-talk."

But the friend talked on, telling the bird about the sun, how it gives life to the world, and how it reigns in the heavens, giving light to each new day. "Look at the sun, Eagle. And when it rises, rise with it. You belong to the sky, not to the earth." At that moment the sun's first rays shot out over the mountain, and suddenly the world was ablaze with light.

The sun rose majestically. The great bird stretched out its wings to greet the sun and feel the warmth on its feathers. The farmer was quiet. The friend said, "You belong not to the earth, but to the sky. Fly, Eagle, fly!" He scrambled back to the farmer. All was silent. The eagle's head stretched up, its wings stretched outwards, and its legs leaned forward as its claws clutched the rock.

Then, without really moving, feeling the updraft of a wind more powerful than any man or bird, the great eagle leaned forward and was swept upward higher and higher, lost to sight in the brightness of the rising sun, never again to live among the chickens.

Questions Fly, Eagle, Fly

1. What did the farmer set out to look for at the beginning of the story?

   (A) a calf

   (B) herders

   (C) rocky cliffs

   (D) an eagle chick

2. Where did the farmer find the eagle chick?

   (A) in its nest

   (B) by the riverbed

   (C) on a ledge of rock

   (D) among the reeds

3. What in the story shows that the farmer was careful with the eagle chick?

   (A) He carried the eagle chick in both hands.

   (B) He brought the eagle chick to his family.

   (C) He put the eagle chick back in its nest.

   (D) He searched the riverbed for the eagle chick.

4. What did the farmer do with the eagle chick when he brought it home?

(A) He taught it to fly.

(B) He set it free.

(C) He trained it to be a chicken.

(D) He made a new nest for it.

5. During the friend's first visit, the eagle chick behaved like a chicken. Give **two** examples that show this.

1 1. _____

_____

1 2. _____

_____

6. When the farmer's friend first met the eagle, how did he try to make the eagle fly?

(A) He lifted it above his head.

(B) He set it on the ground.

(C) He threw it in the air.

(D) He brought it to the mountain.

7. Explain what the farmer's friend meant when he told the eagle, "You belong not to the earth but to the sky."

2

8. Why did the farmer roar with laughter during his friend's first visit?

(A) The eagle was too heavy to fly.

(B) The eagle was difficult to catch.

(C) The eagle looked different from the chickens.

(D) The eagle proved him right.

9. Why did the farmer's friend take the eagle to the high mountains to make it fly? Give **two** reasons.

1 1.

1 2.

10. Find and copy words that tell you how beautiful the sky was at dawn.

1 _____

_____

11. Why was the rising sun important to the story?

(A) It awakened the eagle's instinct to fly.

(B) It reigned in the heavens.

(C) It warmed the eagle's feathers.

(D) It provided light on the mountain paths.

12. You learn what the farmer's friend was like from the things he did.

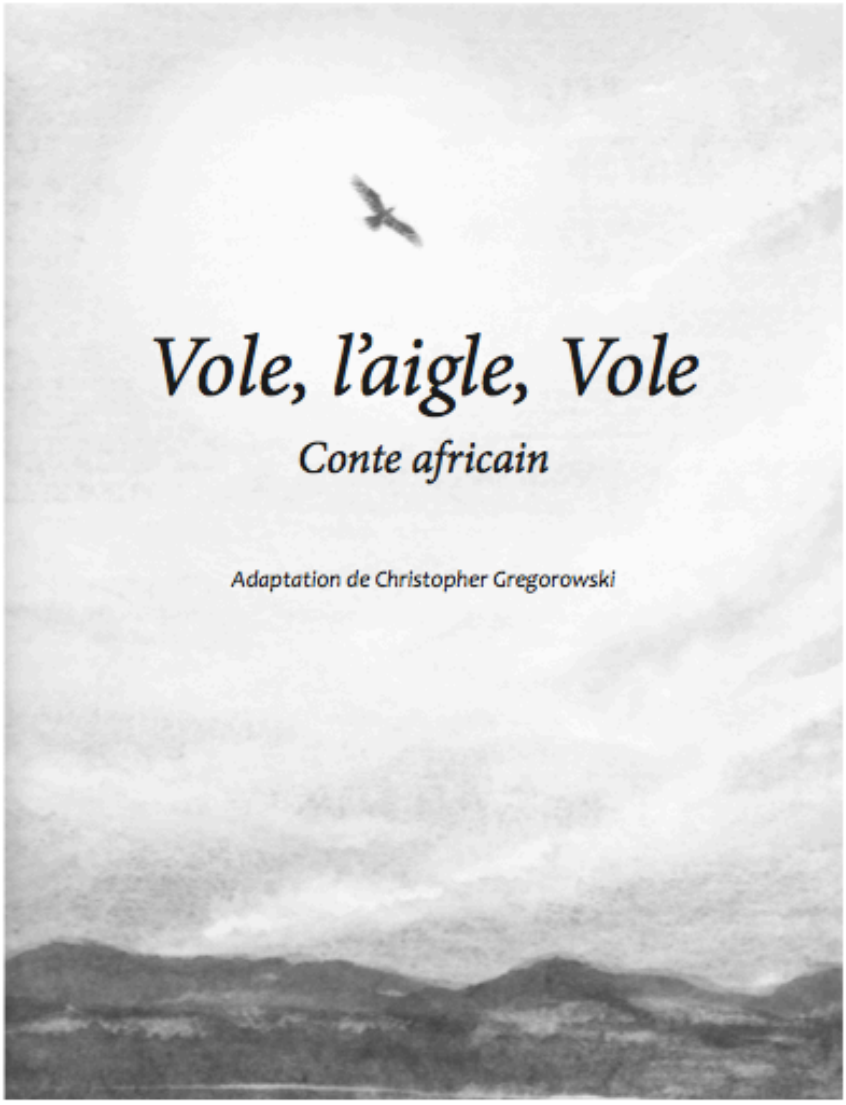
Describe what the friend was like and give an example of what he did that shows this.

2 _____

_____

_____

_____



Vole, l'aigle, Vole

Conte africain

Adaptation de Christopher Gregorowski

Un fermier partit un jour à la recherche d'un veau qui s'était perdu. Les vachers étaient revenus la veille sans lui. Et pendant la nuit, il y avait eu une terrible tempête.

Il se rendit dans la vallée et chercha près de la rivière, parmi les roseaux, derrière les rochers et dans les eaux en mouvement.

Il grimpa les flancs de la haute montagne aux falaises rocheuses. Il regarda derrière un gros rocher au cas où le veau s'y serait caché pour échapper à la tempête. Il n'alla pas plus loin. Là, sur le bord du rocher, il vit quelque chose d'extraordinaire. Un petit aigle, sorti de l'œuf depuis un ou deux jours, avait été jeté hors de son nid par la terrible tempête.

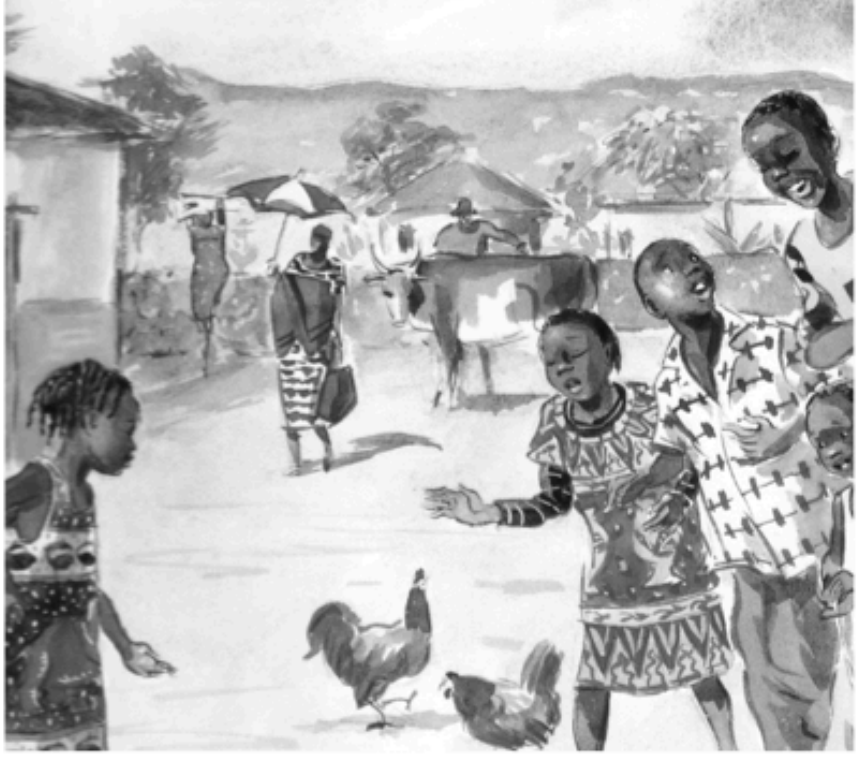
Il s'approcha et prit délicatement le petit oiseau dans ses deux mains. Il allait le ramener chez lui et le soigner.

Il était presque arrivé chez lui quand les enfants coururent à sa rencontre.

« Le veau est revenu tout seul! » crièrent-ils.

Le fermier en fut très heureux. Il montra le petit aigle à sa famille, puis le déposa doucement dans le poulailler au milieu des poules et des poussins.

« L'aigle est le roi des oiseaux, dit-il, mais nous allons lui apprendre à être un poulet. »



Et c'est ainsi que l'aigle vécut parmi les poulets et apprit leurs manières. En grandissant, il ne ressemblait plus guère aux poulets que l'on connaissait.

Un beau jour, un ami en visite aperçut l'oiseau au milieu des poulets.

« Hé, ce n'est pas un poulet! C'est un aigle! »

Le fermier lui dit en souriant : « Bien sûr que c'est un poulet. Regarde! Il marche comme un poulet, il mange comme un poulet. Il pense comme un poulet. Bien sûr que c'est un poulet! »

Mais l'ami ne fut pas convaincu. « Je vais te montrer que c'est un aigle », dit-il.

Les enfants du fermier aidèrent son ami à attraper l'oiseau. Il était assez lourd, mais l'ami du fermier le souleva au-dessus de sa tête et dit : « Tu n'es pas un poulet mais un aigle. Ta place n'est pas sur terre mais dans les airs. Vole, l'aigle, vole! »

L'oiseau étira ses ailes, regarda autour de lui, aperçut les poulets en train de manger et sauta par terre pour chercher de la nourriture avec eux.

« Je t'avais bien dit que c'était un poulet », dit le fermier en éclatant de rire.

Très tôt le lendemain matin, les chiens du fermier se mirent à aboyer. Quelqu'un appelait dans la nuit noire. Le fermier se précipita à la porte. C'était encore son ami. « Laisse-moi une autre chance avec l'oiseau », implora-t-il.

« Sais-tu quelle heure il est? L'aube est encore loin. »

« Viens avec moi. Va chercher l'oiseau. »

À contrecœur, le fermier prit l'oiseau qui dormait profondément au milieu des poulets. Puis, les deux hommes se mirent en route, disparaissant dans la noirceur.

« Où allons-nous? » demanda le fermier d'une voix endormie.

« Dans les montagnes, là où tu as trouvé cet oiseau. »

« Et pourquoi à cette heure ridicule de la nuit? »

« Pour que notre aigle puisse voir le soleil se lever au-dessus de la montagne et qu'il le suive dans les airs, là où se trouve sa place. »

Ils se rendirent dans la vallée et traversèrent la rivière, l'ami marchant en tête. « Dépêche-toi, dit-il, ou l'aube se lèvera avant notre arrivée. »

Le ciel commençait à s'éclaircir tandis qu'ils grimpaient dans la montagne. Les nuages légers, tout d'abord roses, se mirent à briller d'une lumière dorée. Le chemin, parfois dangereux, accroché au flanc de la montagne, leur faisait traverser des passages étroits entre les rochers et les emmenait dans de sombres crevasses d'où ils ressortaient ensuite. « Ici, ça ira », dit-il enfin. Il regarda dans le vide pour apercevoir le sol, des centaines de mètres plus bas. Ils étaient tout près du sommet.

Délicatement, l'ami déposa l'oiseau sur un rocher, de sorte que son regard était tourné vers l'est. Il se mit à lui parler. Le fermier eut un petit rire. « Il ne connaît que la langue des poulets. »

Mais son ami continua de parler à l'oiseau. Il lui parlait du soleil, de sa façon de donner vie au monde et de sa façon de régner sur les cieux, en illuminant chaque jour nouveau. « Regarde le soleil, l'aigle. Et lorsqu'il se lèvera, élève-toi avec lui. Ta place est dans les airs, pas sur terre. » À cet instant, les premiers rayons du soleil pointèrent par-delà la montagne et soudain illuminèrent le monde.

Le soleil se leva majestueusement. Le grand oiseau étira ses ailes pour le saluer et sentir sa chaleur sur son plumage. Le fermier restait silencieux. L'ami dit : « Ta place n'est pas sur terre mais dans les airs. Vole, l'aigle, vole! » Il revint vers le fermier. Tout était calme. L'aigle redressa la tête, étira ses ailes, avança les pattes tandis que ses serres agrippaient le rocher.

Alors, sans véritablement bouger, sentant le souffle d'un vent montant plus puissant que l'homme ou que l'oiseau, le grand aigle se pencha en avant et se laissa emporter vers le ciel de plus en plus haut. Il disparut dans la clarté du soleil levant pour ne plus jamais vivre parmi les poulets.

Questions Vole, l'aigle, vole

1. Que recherche le fermier au début de l'histoire?

   (A) Un veau

   (B) Des vachers

   (C) Des falaises rocheuses

   (D) Un petit aigle

2. Où le fermier trouve-t-il le petit aigle?

   (A) Dans son nid

   (B) Près de la rivière

   (C) Sur le bord d'un rocher

   (D) Parmi les roseaux

3. Dans l'histoire, qu'est-ce qui montre que le fermier fait attention au petit aigle?

   (A) Il porte le petit aigle avec ses deux mains.

   (B) Il amène le petit aigle chez lui.

   (C) Il remet le petit aigle dans son nid.

   (D) Il fouille le lit de la rivière à la recherche du petit aigle.

4. Que fait le fermier avec le petit aigle lorsqu'il l'amène chez lui?

(A) Il lui apprend à voler.

(B) Il lui rend sa liberté.

(C) Il lui apprend à être un poulet.

(D) Il lui fabrique un nouveau nid.

5. Au cours de la première visite de l'ami, le petit aigle s'est comporté comme un poulet. Donne **deux** exemples qui le montrent.

(1) 1. _____

_____

(1) 2. _____

_____

6. Lorsque l'ami du fermier voit l'aigle pour la première fois, comment essaie-t-il de le faire voler?

(A) Il le soulève au-dessus de sa tête.

(B) Il le pose par terre.

(C) Il le jette dans les airs.

(D) Il l'amène dans la montagne.

7. Explique ce que l'ami du fermier veut dire quand il dit à l'aigle : « Ta place n'est pas sur terre mais dans les airs. »

2

8. Pourquoi le fermier éclate-t-il de rire au cours de la première visite?

Ⓐ L'aigle était trop lourd pour voler.

Ⓑ L'aigle était difficile à attraper.

Ⓒ L'aigle paraissait différent des poulets.

Ⓓ L'aigle lui donne raison.

9. Pourquoi l'ami du fermier emmène-t-il l'aigle dans les hautes montagnes pour le faire voler? Donne **deux** raisons.

1 1.

1 2.

10. Trouve et recopie les mots qui décrivent la beauté du ciel à l'aube.

1 _____

_____

11. Pourquoi le lever du soleil joue-t-il un rôle si important dans l'histoire?

(A) Il a réveillé l'instinct de voler de l'aigle.

(B) Il règne sur les cieux.

(C) Il réchauffe le plumage de l'aigle.

(D) Il illumine les chemins montagneux.

12. Le comportement de l'ami du fermier te donne une idée de son caractère.

Décris son caractère et donne un exemple de ce qu'il fait pour justifier ta réponse.

2 _____

_____

_____

_____

The GIANT Tooth Mystery

A fossil is the remains of any creature or plant that lived on the Earth many, many years ago. People have been finding fossils for thousands of years in rocks and cliffs and beside lakes. We now know that some of these fossils were from dinosaurs.

Long ago, people who found huge fossils did not know what they were. Some thought the big bones came from large animals that they had seen or read about, such as hippos or elephants. But some of the bones people found were too big to have come from even the biggest hippo or elephant. These enormous bones led some people to believe in giants.

Hundreds of years ago in France, a man named Bernard Palissy had another idea. He was a famous pottery maker. When he went to make his pots, he found many tiny fossils in the clay. He studied the fossils and wrote that they were the remains of living creatures. This was not a new idea. But Bernard Palissy also wrote that some of these creatures no longer lived on earth. They had completely disappeared. They were extinct.

Was Bernard Palissy rewarded for his discovery? No! He was put in prison for his ideas.

As time went by, some people became more open to new ideas about how the world might have been long ago.

Then, in the 1820s, a huge fossil tooth was found in England. It is thought that Mary Ann Mantell, the wife of fossil expert Gideon Mantell, was out for a walk when she saw what looked like a huge stone tooth. Mary Ann Mantell knew the big tooth was a fossil, and took it home to her husband.

When Gideon Mantell first looked at the fossil tooth, he thought it had belonged to a plant eater because it was flat and had ridges. It was worn down from chewing food. It was almost as big as the tooth of an elephant. But it looked nothing like an elephant's tooth.

Fossil tooth sketched life-sized

Gideon Mantell could tell that the pieces of rock attached to the tooth were very old. He knew that it was the kind of rock where reptile fossils were found. Could the tooth have belonged to a giant, plant-eating reptile that chewed its food? A type of reptile that no longer lived on earth?

Gideon Mantell was really puzzled by the big tooth. No reptile that he knew about chewed its food. Reptiles gulped their food, and so their teeth didn't become worn down. It was a mystery.

Gideon Mantell took the tooth to a museum in London and showed it to other scientists. No one agreed with Gideon Mantell that it might be the tooth of a gigantic reptile.

Gideon Mantell tried to find a reptile that had a tooth that looked like the giant tooth. For a long time, he found nothing. Then one day he met a scientist who was studying iguanas. An iguana is a large plant-eating reptile found in Central and South America. It can grow to be more than two metres long. The scientist showed Gideon Mantell an iguana tooth. At last! Here was the tooth of a living reptile that looked like the mystery tooth. Only the fossil tooth was much, much bigger.

Iguana

A life-sized drawing of an iguana's tooth from Gideon Mantell's notebook

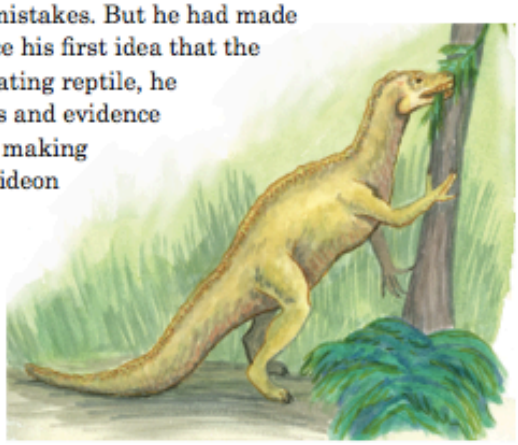
Now Gideon Mantell believed the fossil tooth had belonged to an animal that looked like an iguana. Only it wasn't two metres long. Gideon Mantell believed it was over thirty metres long! He named his creature *Iguanodon*. That means "iguana tooth".

Gideon Mantell did not have a whole *Iguanodon* skeleton. But from the bones he had collected over the years, he tried to figure out what one might have looked like. He thought the bones showed that the creature had walked on all four legs. He thought a pointed bone was a horn. He drew an *Iguanodon* with a horn on its nose.

What Gideon Mantell thought an Iguanodon looked like

Years later, several complete *Iguanodon* skeletons were found. They were only about nine metres long. The bones showed that it walked on its hind legs some of the time. And what Gideon Mantell thought was a horn on its nose was really a spike on its "thumb"! Based on these discoveries, scientists changed their ideas about what the *Iguanodon* looked like.

Gideon Mantell made some mistakes. But he had made an important discovery, too. Since his first idea that the fossil tooth belonged to a plant-eating reptile, he spent many years gathering facts and evidence to prove his ideas were right. By making careful guesses along the way, Gideon Mantell was one of the first people to show that long ago, giant reptiles lived on earth. And then they became extinct.



What scientists today think the Iguanodon looked like

Hundreds of years before, Bernard Palissy had been thrown in prison for saying nearly the same thing. But Gideon Mantell became famous. His discovery made people curious to find out more about these huge reptiles.

In 1842, a scientist named Richard Owen decided that these extinct reptiles needed a name of their own. He called them *Dinosauria*. This means "fearfully great lizard". Today we call them dinosaurs.

1. What is a fossil?

   (A) the surface of rocks and cliffs

   (B) the bones of a giant

   (C) the remains of very old living things

   (D) the teeth of elephants

2. According to the article, why did some people long ago believe in giants?

   1 _____

   _____

   _____

3. Where did Bernard Palissy find fossils?

   (A) on the cliffs

   (B) in the clay

   (C) by a river

   (D) on a path

4. What was Bernard Palissy's new idea?

1 _____

5. Why was Bernard Palissy put into prison?

(A) People were not open to new ideas.

(B) He copied his ideas from Gideon Mantell.

(C) He left tiny fossils in his pottery.

(D) Studying fossils was forbidden in France.

6. Who found the fossil tooth in England?

(A) Bernard Palissy

(B) Mary Ann Mantell

(C) Richard Owen

(D) Gideon Mantell

7. What did Gideon Mantell know about reptiles that made the fossil tooth puzzling?

(A) Reptiles had no teeth.

(B) Reptiles were found under rocks.

(C) Reptiles lived long ago.

(D) Reptiles gulped their food.

8. Gideon Mantell thought the tooth might have belonged to different types of animals. Complete the table to show what made him think this.

Type of animal	What made him think this
A plant eater	The tooth was flat with ridges.
A giant creature	
A reptile	

9. Why did Gideon Mantell take the tooth to a museum?

(A) to ask if the fossil belonged to the museum

(B) to prove that he was a fossil expert

(C) to hear what scientists thought of his idea

(D) to compare the tooth with others in the museum

10. A scientist showed Gideon Mantell an iguana tooth. Why was this important to Gideon Mantell?

1 _____

_____

11. What did Gideon Mantell use when trying to figure out what the *Iguanodon* looked like?

(A) bones he collected

(B) ideas from other scientists

(C) pictures in books

(D) teeth from other reptiles

12. Look at the two pictures of the *Iguanodon*. What do they help you to understand?

2

_____

_____

_____

13. Later discoveries proved that Gideon Mantell was wrong about what the *Iguanodon* looked like. Fill in the blanks to complete the table.

	What Gideon Mantell thought the *Iguanodon* looked like	**What scientists today think the *Iguanodon* looked like**
1	The *Iguanodon* walked on four legs.	
1		The *Iguanodon* had a spike on its thumb.
1	The *Iguanodon* was over 30 metres long.	

14. What were found that showed Gideon was wrong about what the *Iguanodon* looked like?

(A) more fossil teeth

(B) scientific drawings

(C) living *Iguanodons*

(D) whole skeletons

Stop

End of this part of the booklet.
Please stop working.

Le mystère de la dent GÉANTE

Un fossile, ce sont les restes de n'importe quelle créature ou plante qui a vécu sur Terre, il y a beaucoup, beaucoup d'années. On trouve des fossiles depuis des milliers d'années, dans les rochers, les falaises et à côté des lacs. Nous savons maintenant que certains d'entre eux sont des fossiles de dinosaures.

Il y a très longtemps, les gens qui trouvaient des fossiles énormes ne savaient pas ce que c'était. Certains pensaient que les grands os provenaient de gros animaux qu'ils avaient vus ou à propos desquels ils avaient lu quelque chose, tels que les hippopotames ou les éléphants. Mais certains os qui ont été trouvés étaient trop grands pour provenir même des plus gros hippopotames ou éléphants. Ces os énormes ont conduit certaines personnes à croire aux géants.

Il y a des centaines d'années, en France, un homme appelé Bernard Palissy eut une autre idée. C'était un potier célèbre. En allant fabriquer ses pots, il trouva de nombreux petits fossiles dans l'argile. Il étudia les fossiles et écrivit que c'était les restes de créatures vivantes. Ceci n'était pas une idée nouvelle. Mais Bernard Palissy écrivit également que certaines de ces créatures ne vivaient plus sur Terre. Elles avaient complètement disparu. Elles s'étaient éteintes.

Bernard Palissy fut-il récompensé pour sa découverte ? Non ! Il fut mis en prison pour ces idées.

Au fil du temps, certaines personnes sont devenues plus ouvertes aux idées nouvelles qui cherchaient à décrire à quoi le monde avait pu ressembler il y a bien longtemps.

Puis, dans les années 1820, une énorme dent fossile fut découverte en Angleterre. On pense que Mary Ann Mantell, l'épouse de Gideon Mantell, un expert en fossiles, faisait une promenade à pied quand elle aperçut ce qui ressemblait à une énorme dent de pierre. Mary Ann Mantell savait que la grande dent était un fossile et la rapporta à la maison à son mari.

Quand Gideon Mantell jeta un premier coup d'œil à la dent fossile, il pensa qu'elle avait appartenu à un herbivore car elle était plate et striée. Elle était usée parce qu'elle avait servi à mastiquer de la nourriture. Elle était presqu'aussi grosse qu'une dent d'éléphant. Mais elle ne ressemblait pas du tout à une dent d'éléphant.

Dent fossile dessinée grandeur nature

Gideon Mantell réussit à déterminer que les morceaux de rochers accrochés à la dent étaient très vieux. Il savait que c'était le type de rochers dans lesquels on avait trouvé des fossiles de reptiles. La dent pouvait-elle avoir appartenu à un reptile herbivore géant qui mastiquait sa nourriture ? Une sorte de reptile qui ne vivait plus sur Terre ?

Gideon Mantell était très intrigué par la dent géante. Il ne connaissait aucun reptile qui mastiquait sa nourriture. Les reptiles avalaient leur nourriture et leurs dents ne pouvaient donc pas s'user. C'était un mystère.

Gideon Mantell amena la dent à un musée de Londres et la montra à d'autres scientifiques. Personne n'était d'accord avec Gideon Mantell sur le fait que cela pouvait être la dent d'un reptile géant.

Gideon Mantell essaya de trouver un reptile dont la dent ressemblait à la dent géante. Pendant longtemps, il ne trouva rien. Puis un jour, il rencontra un scientifique qui étudiait les iguanes. L'iguane est un grand reptile herbivore qui vit en Amérique centrale et en Amérique du Sud. Il peut atteindre plus de deux mètres de long. Le scientifique montra une dent d'iguane à Gideon Mantell. Enfin ! C'était la dent d'un reptile vivant qui ressemblait à la dent mystérieuse. La seule différence était que la dent fossile était beaucoup, beaucoup plus grande.

Un iguane

Un dessin grandeur nature d'une dent d'iguane, du carnet de Gideon Mantell

Maintenant, Gideon Mantell pensait que la dent fossile avait appartenu à un animal qui ressemblait à un iguane. Seulement, il ne mesurait pas deux mètres. Gideon Mantell pensait qu'il mesurait plus de trente mètres de long ! Il appela sa créature *Iguanodon*. Ce qui signifie « dent d'iguane ».

Gideon Mantell n'avait pas un squelette d'*Iguanodon* complet. À partir des os qu'il avait récoltés au fil des ans, il essaya d'imaginer l'apparence extérieure de celui-ci. Il pensa que les os montraient que la créature avait marché sur ses quatre pattes. Il supposa qu'un os pointu était une corne. Il dessina un *Iguanodon* avec une corne sur le nez.

L'apparence extérieure d'un Iguanodon d'après Gideon Mantell à cette époque-là

Des années plus tard, plusieurs squelettes complets de l'*Iguanodon* ont été trouvés. Ceux-ci ne mesuraient que neuf mètres de long. Les os montraient qu'il lui arrivait de marcher sur ses pattes arrière. Et ce que Gideon Mantell prenait pour une corne sur son nez était en réalité un piquant sur son « pouce » ! En se fondant sur ces découvertes, les scientifiques ont changé d'idée sur ce à quoi ressemblait l'*Iguanodon*.

Gideon Mantell a fait quelques erreurs mais il a également fait une découverte importante. Depuis sa première idée que la dent fossile appartenait à un reptile herbivore, il a passé de nombreuses années à récolter des faits et des preuves pour montrer que ses idées étaient exactes. Grâce à ses suppositions prudentes, Gideon Mantell a été l'une des premières personnes à montrer qu'il y a longtemps des reptiles géants vivaient sur terre. Puis, ils se sont éteints.

L'apparence extérieure d'un Iguanodon d'après les scientifiques aujourd'hui

Des centaines d'années auparavant, Bernard Palissy avait été jeté en prison pour avoir dit presque la même chose. Mais Gideon Mantell est devenu célèbre. Sa découverte a rendu les gens curieux de découvrir d'autres choses sur ces énormes reptiles.

En 1842, un scientifique du nom de Richard Owen décida que ces reptiles disparus devaient avoir leur propre nom. Il les appela *Dinosauria*, ce qui signifie « lézard terriblement grand ». Aujourd'hui, nous les appelons dinosaures.

Questions Le mystère de la dent géante

1. Qu'est-ce qu'un fossile ?

   (A) La surface de rochers ou de falaises

   (B) les os d'un géant

   (C) les restes de choses vivantes très anciennes

   (D) les dents d'un éléphant

2. Selon l'article, pourquoi certaines personnes croyaient-elles il y a très longtemps aux géants ?

3. Où Bernard Palissy a-t-il trouvé des fossiles ?

   (A) sur les falaises

   (B) dans l'argile

   (C) près d'une rivière

   (D) sur un chemin

4. Quelle a été la nouvelle idée de Bernard Palissy ?

_____

5. Pourquoi a-t-on mis Bernard Palissy en prison ?

(A) Les gens n'étaient pas ouverts à ses idées nouvelles.

(B) Il avait copié ses idées sur celles de Gideon Mantell.

(C) Il avait laissé des petits fossiles dans sa poterie.

(D) En France, il était interdit d'étudier les fossiles.

6. Qui a trouvé la dent fossile en Angleterre ?

(A) Bernard Palissy

(B) Mary Ann Mantell

(C) Richard Owen

(D) Gideon Mantell

7. Que savait Gideon Mantell sur les reptiles qui lui a fait comprendre le caractère intriguant de la dent fossile ?

(A) Les reptiles n'avaient pas de dents.

(B) Les reptiles ont été trouvés sous les rochers.

(C) Les reptiles vivaient il y a longtemps.

(D) Les reptiles avalaient leur nourriture.

8. Gideon Mantell pensait que la dent avait pu appartenir à différentes sortes d'animaux. Complète le tableau pour montrer ce qui lui faisait penser cela.

Sorte d'animal	**Ce qui lui faisait penser cela**
Un herbivore	La dent était plate et striée.
Une créature géante	
Un reptile	

9. Pourquoi Gideon Mantell a-t-il amené la dent à un musée ?

(A) pour demander si le fossile appartenait au musée

(B) pour prouver qu'il était expert en fossiles

(C) pour écouter ce que les scientifiques pensaient de son idée

(D) pour comparer la dent à d'autres dans le musée

10. Un scientifique a montré une dent d'iguane à Gideon Mantell. Pourquoi cela a-t-il été important pour Gideon Mantell ?

1 _____

_____

11. Qu'est-ce que Gideon Mantell a utilisé quand il a essayé d'imaginer l'apparence extérieure d'un *Iguanodon* ?

(A) les os qu'il avait récoltés

(B) les idées d'autres scientifiques

(C) des images dans des livres

(D) les dents d'autres reptiles

12. Regarde les deux images de l'*Iguanodon*. Que te permettent-elles de comprendre ?

(2)

_____

_____

_____

13. Les découvertes suivantes ont prouvé que Gideon Mantell avait tort concernant l'apparence extérieure de l'*Iguanodon*. Remplis les espaces pour compléter le tableau.

	L'apparence extérieure de l'*Iguanodon* d'après Gideon Mantell à cette époque-là	**L'apparence extérieure de l'*Iguanodon* d'après les scientifiques aujourd'hui**
(1)	L'*Iguanodon* marchait sur ses quatre pattes.	
(1)		L'*Iguanodon* avait un piquant sur le pouce.
(1)	L'*Iguanodon* mesurait 30 mètres de long.	

14. Qu'a-t-on trouvé qui montre que Gideon avait tort concernant l'apparence extérieure de l'*Iguanodon* ?

(A) plus de dents fossiles

(B) des dessins scientifiques

(C) des *Iguanodons* vivants

(D) des squelettes entiers

Arrête

Fin de cette partie du cahier.
S'il te plaît, arrête d'écrire.

Appendix B: Instructions and criteria for expert reviewers

INSTRUCTIONS:

1. Review the following text and items and rate according to the criteria below:

Rating Criteria

(0)	**No difference**	No difference in meaning between the two versions
(1)	**Negligible difference**	Minimal differences in meaning between the two versions
(2)	**Somewhat different**	There are clear differences in meaning between the two versions but they are not expected to lead to differences in performance between the two groups of examinees
(3)	**Very different**	There are clear differences in meaning between the two versions that are expected to lead to differences in performances between the two groups of examinees.

2. Please determine how confident you are when giving your rating an item.
   (0) Not confident (1) Somewhat confident (2) Confident (3) Very confident

3. If there is a difference, please determine the language group (French and English) the item would favor.

4. Please indicate the code for the source(s) of difference. Refer to the 'Codes for the sources of translation difference' sheet.

5. Please describe the translation problems.

Adapted from K.Y. Jang (2010)

Appendix C: Checklist of Possible Linguistic, Cultural and Format Differences

The following checklist is a summary of the types of differences that can be found between different language versions of a test. As you review each passage and item, please consider the checklist of differences. This review sheet was adopted from Ercikan et al., 2013.

Code No.	Source of Differences	Review Questions & Examples
1	**Differences in words, expressions and structure of sentence inherent to a language or culture**	**Are there differences in words, expressions or sentence structure that are inherent in one language or culture?** Example: The English sentence "Most roller bladers do not favour a helmet bylaw" was translated into French with the expressions ""la plupart des personnes qui font du patin à roulettes " and "un règlement municipal du casque protecteur" for "helmet bylaw." These language versions are different because the meaning of the French words is different.
2	**Differences in word difficulty or familiarity of vocabulary**	**Are their particular words or vocabulary that are more commonly used or less formal in one language than the other?** Example: In the English version examinees are asked, "What linear equation defines the total cost *C* for any number *n* of T-shirts?" Translated into French as the expression, "Quelle équation du premier degré définit le coût total *C* en fonction du nombre *n* de t-shirts? (Ercikan et al., 2010) The English sentence is more direct or concrete.
3	**Differences in cohesiveness and continuity of text**	**Are there grammatical and lexical differences that make one language version more coherent and understandable?**

Code No.	Source of Differences	Review Questions & Examples
		Example: Repetition of key words in the English version help to make the structure of the text clearer.
4	**Differences in meaning**	**Does the meaning of a word differ because it has multiple meanings in one language?** Example: In English there are several meanings for the word "cry". An item in the English version of a test states, "Sometime in the night the *cry* awoke her, a sound so anguished she was on her feet before she was awake." The Finnish version of the test states, "Johonkin aikaan yöstä *kiljahdus* herätti hänet. Ääni oli niin tuskainen, että hän oli jalkeilla ennen kuin ehti herätäkään." The English translation of the above Finnish version reads as follows: "Sometime in the night a *cry/shout* awoke her. The sound was so anguished that she was on her feet even before she had had time to awake" (Arffman, 2007).
5	**Omissions or additions of words or phrases that affect meaning**	**Does one language version omit or include words that influence the meaning of a word or item?** Example: The English item included the expression "this number written in standard form" and the French item had the phrase "ce nombre est" ("this number is")." The idea of "standard form" was excluded from the French translation (Oliveri, 2008).
6	**Differences in verb tense**	**Does the verb tense differ in one language version?**

Code No.	Source of Differences	Review Questions & Examples
		Example: In the English version the word "reads" (present tense) was written as "a lu" (past tense) in the French version (Oliveri, 2008).
7	**Differences in length or sentence complexity that make the item more difficult for one language group**	**Are there differences in length or complexity of sentences that are likely to affect the performance of one linguistic group?** Example: Is it harder to understand a sentence in one language version because it is longer or the content is more difficult?
8	**Differences in additional information that guides how examinees' think**	**Does one language version include prompts or information that guide examinees' thinking?** Example: A question asked in the English version, "At what point will the reflection of the candle *appear to be?"* The question in the French version was, "At what point will the image of the candle *seem to appear to be?"* The French version is more informative, as it suggests that the reflection in the mirror may seem different than the actual object.
9	**Differences in item format**	**Are there format differences such as with punctuation, capitalization, item structure or distractors?** Example: A word was presented only in the stem of an English item but it was presented in all four options in the French item.
10	**Differences in cultural relevance**	**Is a passage or item more relevant to one group than the other?**

Code No.	Source of Differences	Review Questions & Examples
		Example: A passage contains cultural events, conventions or references that are more familiar to one group.
11	**Differences due to inappropriate translation**	**Are there any key words that were inappropriately translated?** Example: In the English version of a text the sentence is, "How many days, *she* wondered, had she sat like this, watching…" In the Finnish version of the text the sentence is, Montakohan päivää, *nainen* mietti, hän olikaan jo istunut tällä tavoin katsellen..." Translated in English the Finnish version states, "How many days, *woman* wondered, had she sat like this, watching…(Arffman, 2010).
12	**Differences in reading processes assessed by the two language versions**	**Does an item assess the same reading processes in both the French and English language versions?** Example: The item for the French version asks about the central ideas in the passage. For example, "Que recherché le fermier au début de l'histoire?"(Foy & Drucker, 2013). The item for the English version asks examinees to interpret information. The question was, "Explain what the farmer's friend meant when he told the eagle you belong not to the earth but the sky" (Foy & Drucker, 2013).
13	**Other**	**Describe a difference unaccounted for by other codes.**

Appendix D: Item p-Values for Booklets 1-13

Item p-values for booklets 1 and 2

	Booklet 1		Booklet 2	
Item	French	English	French	English
1	.813	.880	.904	.935
2	.833	.813	.809	.762
3	.787	.885	.503	.716
4	.887	.909	.881	.909
5	.762	.692	.587	.779
6	.570	.690	.367	.756
7	.476	.585	.472	.677
8	.737	.784	.171	.567
9	.636	.866	.308	.439
10	.882	.954	.404	.735
11	.611	.702	.559	.638
12	.219	.469	.394	.623
13	.555	.922	.557	.688
14	.877	.926	.727	.764
15	.730	.775	.317	.567
16	.474	.738	.340	.688
17	.893	.906	.316	.691
18	.835	.745	.328	.688
19	.660	.728	.428	.798

	Booklet 1		Booklet 2	
Item	French	English	French	English
22	.410	.487	.231	.473
23	.575	.720	.328	.667
24	.569	.698	.562	.716
25	.477	.648	.495	.550
26			.596	.592

Item p-values for booklets 3 and 4

	Booklet 3		Booklet 4	
Item	French	English	French	English
1	.526	.667	.878	.911
2	.738	.743	.873	.859
3	.492	.529	.714	.708
4	.683	.685	.568	.404
5	.603	.691	.671	.744
6	.620	.653	.643	.664
7	.869	.820	.714	.855
8	.671	.789	.840	.845
9	.780	.811	.570	.628
10	.263	.551	.362	.485
11	.672	.701	.751	.837
12	.604	.707	.477	.628

	Booklet 3		Booklet 4	
Item	French	English	French	English
13	.459	.495	.859	.909
14	.599	.539	.730	.773
15	.825	.876	.648	.600
16	.868	.887	.533	.604
17	.774	.657	.516	.682
18	.590	.418	.573	.714
19	.651	.778	.439	.678
20	.626	.589	.538	.654
21	.754	.855	.887	.835
22	.811	.798	.439	.535
23	.563	.591	.523	.628
24	.334	.417	.357	.402
25	.692	.787	.425	.793
26	.539	.629	.575	.634
27	.856	.915	.587	.610
28	.749	.773	.387	.420
29	.677	.615	.735	.759
30	.563	.628	.298	.429
31	.385	.590	.707	.799
32	.585	.728	.299	.481
33	.517	.720		

	Booklet 3		Booklet 4	
Item	French	English	French	English
34	.551	.676		

Item p-values for booklets 5 and 6

	Booklet 5		Booklet 6	
Item	French	English	French	English
1	.878	.813	.790	.844
2	.873	.542	.658	.576
3	.714	.702	.169	.316
4	.568	.415	.583	.636
5	.671	.728	.872	.925
6	.643	.677	.776	.832
7	.714	.583	.771	.624
8	.840	.479	.660	.663
9	.849	.777	.682	.844
10	.362	.449	.614	.707
11	.751	.813	.507	.692
12	.477	.545	.498	.601
13	.859	.870	.742	.681
14	.730	.600	.795	.802
15	.648	.353	.704	.622
16	.533	.600	.549	.596

	Booklet 5		Booklet 6	
Item	French	English	French	English
17	.397	.949	.364	.543
18	.573	.872	.651	.713
19	.439	.615	.476	.522
20	.538	.644	.504	.584
21	.887	.835	.646	.729
22	.439	.695	.290	.403
23	.523	.749	.636	.620
24	.446	.603	.547	.590

Item p-values for booklets 7 and 8

	Booklet 7		Booklet 8	
Item	French	English	French	English
1	.793	.755	.795	.868
2	.834	.828	.852	.863
3	.706	.633	.808	.861
4	.595	.563	.852	.865
5	.419	.617	.772	.696
6	.695	.719	.540	.693
7	.587	.634	.494	.576
8	.533	.581	.718	.747
9	.660	.792	.674	.842

	Booklet 7		Booklet 8	
Item	French	English	French	English
10	.293	.443	.899	.942
11	.692	.635	.595	.699
12	.629	.634	.220	.484
13	.686	.770	.552	.869
14	.487	.550	.707	.780
15	.547	.659	.514	.595
16	.622	.793	.614	.734
17	.383	.515	.691	.804
18	.764	.820	.347	.531
19	.339	.469	.829	.855
20	.281	.530	.333	.349
21	.425	.425	.321	.520
22	.358	.482	.426	.387
23	.726	.795	.446	.480
24	.433	.501	.676	.737
25	.640	.612	.510	.492
26	.283	.418	.631	.660
27	.358	.543	.327	.411
28	.337	.429	.398	.515
29			.342	.462

Item p-values for booklets 9 and 10

	Booklet 9		Booklet 10	
Item	French	English	French	English
1	.778	.836	.883	.940
2	.851	.828	.778	.800
3	.783	.868	.492	.738
4	.893	.903	.904	.928
5	.740	.717	.821	.770
6	.569	.670	.724	.767
7	.528	.581	.658	.659
8	.785	.763	.355	.620
9	.667	.859	.413	.464
10	.916	.927	.580	.736
11	.548	.704	.599	.684
12	.177	.496	.444	.549
13	.551	.901	.805	.866
14	.852	.802	.655	.626
15	.439	.493	.171	.346
16	.553	.627	.633	.682
17	.401	.398	.923	.954
18	.398	.734	.816	.865
19	.557	.633	.754	.640
20	.565	.583	.676	.671

	Booklet 9		Booklet 10	
Item	French	English	French	English
21	.400	.404	.420	.089
22	.745	.750	.668	.689
23	.320	.404	.540	.717
24	.734	.783	.513	.577
25	.361	.556		

Item p-values for booklets 11 and 12

	Booklet 11		Booklet 12	
Item	French	English	French	English
1	.546	.672	.850	.895
2	.743	.796	.848	.890
3	.480	.568	.683	.666
4	.683	.722	.587	.346
5	.654	.707	.655	.752
6	.612	.690	.546	.657
7	.858	.830	.710	.854
8	.735	.809	.831	.858
9	.795	.853	.564	.600
10	.306	.563	.355	.406
11	.710	.748	.729	.832
12	.582	.703	.469	.656

	Booklet 11		Booklet 12	
Item	French	English	French	English
13	.501	.568	.845	.897
14	.578	.630	.708	.787
15	.752	.686	.647	.657
16	.830	.792	.567	.645
17	.710	.679	.440	.500
18	.499	.585	.597	.706
19	.321	.526	.499	.677
20	.643	.704	.497	.637
21	.500	.604	.697	.774
22	.577	.569	.480	.597
23	.629	.771	.608	.704
24	.321	.428	.706	.811
25	.667	.652	.347	.496
26	.664	.651	.768	.832
27			.372	.387
28			.322	.493
29			.391	.426
30			.391	.522
31			.711	.744
32			.493	.514
33			.654	.654

	Booklet 11		Booklet 12	
Item	French	English	French	English
34			.274	.436
35			.407	.487
36			.376	.410

Item p-values for booklet 13

	Booklet 13	
Item	French	English
1	.812	.849
2	.797	.847
3	.784	.831
4	.419	.553
5	.506	.602
6	.778	.836
7	.551	.684
8	.727	.836
9	.773	.888
10	.824	.921
11	.737	.801
12	.557	.747
13	.938	.887
14	.522	.716

	Booklet 13	
Item	French	English
15	.430	.629
16	.370	.492
17	.709	.832
18	.397	.638
19	.649	.794
20	.196	.300
21	.540	.604
22	.745	.747
23	.418	.695
24	.335	.430
25	.478	.560
26	.213	.321
27	.591	.593
28	.476	.512
29	.542	.595
30	.230	.250
31	.585	.635
32	.646	.681
33	.631	.712
34	.546	.548
35	.506	.611

Appendix E: Items Identified as DIF by LR and OLR methods

Booklet 1

Item	χ^2 (dif)	df	p value	R^2	DIF Type
5	35.68	2	0.00	.049*	Uniform
9	42.47	2	0.00	.040*	Uniform
12	45.58	2	0.00	.046*	Uniform
13	131.29	2	0.00	.159**	Uniform
16	53.70	2	0.00	.073**	Uniform
18	59.27	2	0.00	.068*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 2

Item	χ^2 (dif)	df	p value	R^2	DIF Type
3	38.06	2	0.00	.052*	Uniform
5	54.13	2	0.00	.064*	Uniform
8	37.03	2	0.00	.042*	Uniform
19	22.72	2	0.00	.036*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 3

Item	χ^2 (dif)	df	p value	R^2	DIF Type
7	21.71	2	0.00	.038*	Both
10	62.09	2	0.00	.080**	Uniform
17	45.55	2	0.00	.055*	Uniform
18	50.26	2	0.00	.064*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large

effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 4

Item	χ^2 (dif)	df	p value	R^2	DIF Type
4	53.32	2	0.00	.070**	Uniform
25	107.50	2	0.00	.124**	Uniform
32	56.51	2	0.00	.058*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 5

Item	χ^2 (dif)	df	p value	R^2	DIF Type
1	33.37	2	0.00	.089**	Uniform
5	63.10	2	0.00	.080**	Uniform
15	56.59	2	0.00	.060*	Uniform
19	67.42	2	0.00	.090**	Uniform
23	57.33	2	0.00	.050*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 6

Item	χ^2 (dif)	df	p value	R^2	DIF Type
3	39.68	2	0.00	.040*	Uniform
7	51.60	2	0.00	.070**	Uniform
9	26.76	2	0.00	.040*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 8

Item	χ^2 (dif)	df	p value	R^2	DIF Type
5	25.72	2	0.00	.037*	Uniform
12	59.43	2	0.00	.062*	Uniform
13	84.93	2	0.00	.104**	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 9

Item	χ^2 (dif)	df	p value	R^2	DIF Type
12	100.61	2	0.00	.101*	Uniform
13	115.02	2	0.00	.140**	Both
18	83.49	2	0.00	.096*	Both

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 10

Item	χ^2 (dif)	df	p value	R^2	DIF Type
3	51.76	2	.000	.065*	Uniform
5	39.91	2	.000	.042*	Uniform
8	42.51	2	.000	.048*	Uniform
15	53.77	2	.000	.053*	Uniform
19	48.01	2	.000	.058*	Uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. Two asterisks indicate large effect sizes, $R^2 \geq 0.070$ and one asterisk indicates moderate effect size, $R^2 > 0.035$ and < 0.070.

Booklet 13

Item	χ^2 (dif)	df	p value	R^2	DIF Type
1	1205.63	2	.000	.672*	Uniform
7	53.41	2	.000	.041*	Uniform and non-uniform
13	1253.64	2	.000	.665*	Uniform
23	54.41	2	.000	.041*	Uniform and non-uniform

Note. χ^2 (dif) = Chi-square difference. df = degrees of freedom. One asterisk indicates moderate effect size. R^2 =