frontiers in Education

ORIGINAL RESEARCH
published: 06 December 2019
doi: 10.3389/feduc.2019.00120

Measurement Comparability of Reading in the English and French Canadian Populations: Special Case of the 2011 Progress in International Reading Literacy Study

Shawna Goodrich[1]* *and Kadriye Ercikan*[2]*

[1] Department of National Defence, Ottawa, ON, Canada, [2] Educational Testing Service, University of British Columbia, Princeton, NJ, United States

The purpose of this study is to examine item equivalence and score comparability of the Progress in International Reading Literacy Study (PIRLS) 2011 for the Canadian French and English language groups. Two methods of differential item functioning were conducted to examine item equivalence across 13 test booklets designed to assess reading literacy in early years of schooling. Four bilingual reviewers with expertise in reading literacy conducted independent, linguistic, and cultural reviews to identify both the degree of item equivalence and potential sources of differences between language versions of released items. Results indicate that an average of 25% of items per booklet function differentially at the item level. Reviews by experts indicate differences between the two language versions on some items flagged as displaying differential item functioning (DIF). Some of these were identified to have linguistic differences pointing to differential difficulty levels in the two language versions.

Keywords: score comparability, test equivalence, item equivalence, differential item functioning, international assessment, reading literacy achievement

OPEN ACCESS

Edited by:
Elizabeth Archer,
University of the Western Cape,
South Africa

Reviewed by:
Vanessa Scherman,
University of South Africa, South Africa
Jeffrey K. Smith,
University of Otago, New Zealand

***Correspondence:**
Shawna Goodrich
shawnago@gmail.com
Kadriye Ercikan
kercikan@ets.org

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *26 July 2019*
Accepted: *09 October 2019*
Published: *06 December 2019*

Citation:
Goodrich S and Ercikan K (2019) Measurement Comparability of Reading in the English and French Canadian Populations: Special Case of the 2011 Progress in International Reading Literacy Study. Front. Educ. 4:120. doi: 10.3389/feduc.2019.00120

The Progress in International Reading Literacy Study (PIRLS) 2011 assessments were 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 such large-scale assessments (LSAs) serve two general purposes. First, within and across countries, results are used to draw comparative conclusions across groups about academic achievement, learning, and educational accountability (Crundwell, 2005; Johansson, 2016; Klinger et al., 2016). Second, they are intended to provide evidence to policymakers and administrators about the strengths and weaknesses of educational systems. Often, evidence from LSAs have been used to inform decisions about the structure and delivery of education within nations and has spurred reforms for educational programs (Canadian Education Statistics Council, 2016; Tobin et al., 2016). Given the influential role of comparisons based on international LSAs on educational policy worldwide, it is essential to ensure that tests are equivalent across linguistic and cultural groups in order for decisions and consequences based on results to be fair, justified, and constructive. 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 multiple 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 (Fairbairn and Fox, 2009; Rogers et al., 2010; Marotta et al., 2015). Yet, research

conducted in Canada comparing the French and English versions of LSAs has found that 18–60% of items function differentially for the two groups (Gierl et al., 1999; Gierl, 2000; Ercikan and McCreith, 2002; Ercikan et al., 2004b; Oliveri and Ercikan, 2011; Marotta et al., 2015).

Non-equivalence across different language test versions may be attributable to a range of factors that include cultural and curricular differences; however, differences due to test translation play a major role in such equivalence (Bachman, 2000; Cummins, 2000; Tanzer, 2005; Cohen, 2007; Solano-Flores et al., 2012). This is due in part to the inherent differences between languages. Languages differ in both form and meaning (Steiner and Mahn, 1996). Form relates to sentence structure, writing systems, word order, ways of conveying new information, ways of signaling thematic structures, and methods of cohesion (Baker, 1992; Arffman, 2007, 2010; Grisay and Monseur, 2007; He and van de Vijver, 2012). Meaning is inextricably interrelated with language (Rorty, 1977). Human beings acquire and use language to explore and create meaning by engaging in interpersonal relationships and interpreting experiences (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 are interwoven with social and cultural practices (Boroditsky, 2011; Roth et al., 2013). Previous research has identified a number of differences between languages that make test adaptation difficult and can result in the incomparability of assessment scores between language groups (Ercikan, 1998; Bonnet, 2002; Arffman, 2007; Grisay and Monseur, 2007; Sireci, 2008; Marotta et al., 2015). These include differences in grammar, meaning, vocabulary, syntax, word usage, and difficulty (Allalouf and Sireci, 1998; Ercikan, 1998; Allalouf et al., 1999; Ercikan et al., 2004a; Arffman, 2007).

Research has also demonstrated that psychometric differences between language versions may be attributable to factors other than test adaptation, such as cultural and curricular differences between groups (Solano-Flores and Nelson-Barber, 2000, 2001; Ercikan et al., 2004a). Examinees not only draw upon their language to make sense of words and texts but also diverse social and cultural experiences that create different linguistic repertoires (Steiner and Mahn, 1996; Greenfield, 1997; Gee, 2001; Arffman, 2007). Words are not inherently meaningful; social and cultural interactions, and conventions imbue words with meaning (Greenfield, 1997; Derrida, 1998; Campbell and Hale, 2003). Cultural experiences may produce different interpretations of commonly shared words, which affect the trajectory of thought processes and ultimately responses to test questions (Solano-Flores, 2006; Roth, 2009; Ercikan and Lyons-Thomas, 2013).

PURPOSE

The focus of this study is the international LSA program that assesses reading literacy in early years of schooling in Canada. PIRLS is administered in Canada to students in 4th grade to determine achievement levels of students at provincial and national levels. To examine the equivalence of French and English language versions of PIRLS 2011, two DIF detection methods were used and expert reviews were conducted. As access to all 10 passages and items from PIRLS 2011 was unavailable, expert reviews were restricted to the four reading passages and accompanying items released to the public.

METHOD

In 2011, PIRLS was administered to students in their fourth year of formal schooling in 48 countries (International Association for the Evaluation of Educational Achievement, 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 taking the test in English and 6,500 in French. Nine Canadian provinces participated: British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, Nova Scotia, and Newfoundland, and Labrador. In addition to assessing reading achievement, background information regarding students, home supports for literacy, teachers, schools, and curriculum is collected.

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 13–16 questions or items pertaining to the passage. Each student completes one randomly assigned booklet that contains two blocks. A total of five literary and five informational passages or blocks were distributed across individual booklets (International Association for the Evaluation of Educational Achievement, 2012). Reading achievement data from PIRLS is used to create a scale with a mean set to 500 and standard deviation to 100.

Reading literacy is defined by PIRLS as "the ability to understand and use those written language forms required by society and/or valued by the individual" (Mullis et al., 2009, p. 11). PIRLS focuses on two reading literacy purposes, which include reading for literary experience and reading to acquire and use information (Labrecque et al., 2012). Four processes of comprehension that relate to how readers construct meaning from text are targeted by PIRLS. Each item is designed to measure one of the following reading comprehension processes: (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 (International Association for the Evaluation of Educational Achievement, 2012).

The French language version of PIRLS is developed through the forward translation of the English language version (International Association for the Evaluation of Educational Achievement, 2012). With the forward translation method, a monolingual test developer constructs the test in a source language. Bilingual translators then adapt the test from the source language to the target language. Bilingual translators check the equivalence of the two tests. Guidelines created by the IEA to assist in the translation of all the assessment materials stipulate that each participating country is responsible

for adapting PIRLS material to their cultural context. 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. The guidelines provide recommendations that include the preservation of the original information, the use of correct grammar and punctuation, and preserving the meaning of idiomatic expressions rather than literal adaptations.

Differential Item Functioning Analysis

Identification of items with DIF across different linguistic versions of LSAs indicates that these items may not function the same way across linguistic groups for examinees of 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 different expected probabilities of answering an item correctly, as indicated by DIF, the item parameters for these items are not comparable and may compromise the equivalence of measurement for the two comparison groups.

The measurement model that is used in PIRLS is based on item response theory (IRT). With IRT modeling of responses, the discrimination and difficulty parameters characterize the most important aspects of measurement equivalence with achievement tests. The discrimination parameter indicates how rapidly the item characteristic curve (ICC) rises at the point of inflection, representing the degree to which an item response varies by ability level. The difficulty parameter indicates the location of the item on the ability scale in which examinees have a 0.5 probability level of answering the item correctly. Items are typically flagged for DIF if response probabilities for examinees at the same ability levels depend on group membership. As different methods for identifying DIF may not give identical results, the use of more than one method is recommended, to allow for the corroboration of DIF status for the items analyzed (Ercikan and McCreith, 2002; Oliveri and Ercikan, 2011). In this research, an IRT-based approach and logistic/ordinal logistic regression approaches were used.

IRT-Based DIF

An application of the Linn and Harnisch (LH) method (Linn and Harnisch, 1981) to IRT-based item parameters 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 data from each respective group. IRT parameters were calibrated using PARDUX software (CTB/McGraw-Hill, 1991). PARDUX software uses marginal maximum likelihood procedures to simultaneously generate item parameters for dichotomous and polytomous items. This software was 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, a statistical test used to determine whether two means differ significantly. 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 $|p_{diff}| \leq 0.10$ are identified as moderate magnitude DIF, Level 2. Large magnitude DIF, Level 3, is identified by $|Z| \geq 2.58$ and $|p_{diff}| \geq 0.10$ (Yen, 1993; Ercikan and McCreith, 2002). A three-parameter logistic model (Lord, 1980) was used to calibrate multiple-choice items and the two-parameter partial credit (2PPC) model (Yen, 1993) was used to calibrate constructed-response items. The 2PPC is a special case of Bock's nominal model and is equivalent to the generalized partial credit model (Muraki, 1992; Yen and Fitzpatrick, 2006). This model allows polytomous items to vary in their discrimination. To examine the fit of a unidimensional model with the data, the assumptions of local independence and model fit were examined using the Q_3 and Q_1 statistics (Yen, 1993), respectively.

Logistic Regression/Ordinal Logistic Regression

The logistic regression (LR) DIF method (Swaminathan and Rogers, 1990) 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. The ordinal logistic regression (OLR) DIF method (Miller and Spray, 1993) 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 falling within or below thresholds across the number of response categories minus one. The LR and OLR methods can be used for binary and ordinal response data, respectively. These methods also provide a means to flag items as either uniform or non-uniform DIF. 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 above the levels of the latent trait.

Zumbo (1999) recommends conducting both the Chi-square test and a measure of effect size to identify DIF with LR and to ensure that the 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 was *R*-squared, which indicates the proportion of shared variance between two or more variables. Items are classified with DIF 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 $0.035 \geq R^2 \leq .070$, 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. For simplicity, effect sizes in this study are reported according to magnitude rather than numerical values. When DIF statistical conclusions are based on both the *p*-value for the Chi-square difference test and the effect size criterion with LR and OLR methods, the Type I error rate and statistical power is conservative (Zumbo, 2008).

Bilingual Expert Reviews

Four bilingual reviewers with expertise in reading literacy and experience with test construction evaluated the equivalence of released PIRLS 2011 items in French and English. 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 the other two reviewers. Two of the experts have extensive experience with bilingual test development, particularly across French and English language versions, and have expertise in reading literacy. The other two reviewers have elementary and middle school teaching experience. Experts examined items in the two language versions considering cultural relevance, and equivalence of meaning, overall format, inadvertent cues given to examinees to solve the problems, and the maintenance of intended reading and difficulty levels across test versions (Bowles and Stansfield, 2008; Ercikan and Lyons-Thomas, 2013).

A training session informed by previous research (Gierl and Khaliq, 2001; Ercikan, 2003) was conducted to ensure that reviewers understood the specific adaptation problems associated with the linguistic and cultural 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. Reviewers were given copies of the four passages in French and English, instructions, and criteria for rating the significance of differences between test language versions; a checklist of potential translation errors; and worksheets to code errors. 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 and Lyons-Thomas, 2013). The group was 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, 2003). 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. In the first stage, all items were reviewed independently without knowledge about which items were identified as DIF. In the second stage, the group only reviewed items that were identified as DIF for which there were rating differences among reviewers. Each reviewer stated how they had rated an item, described the differences identified, and explained their rationale for a rating. Although the purpose of this study did not include an examination of differences across passages, reviewers noted potential problems within passages and differences that may impact student performance on associated items. Once each reviewer had an opportunity to explain their ratings, a group discussion about the nature and degree of differences with respect to the rating criteria followed. The discussion continued until they reached a consensus about a rating.

TABLE 1 | Descriptive statistics for booklets 1–13.

		French		English		
Booklet	**No. 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 (7.50)	500	23.12 (7.75)	2.34*
12	36	433	23.02 (9.30)	500	26.84 (9.19)	3.82*
13	35	681	19.64 (8.95)	731	25.71 (9.51)	6.07*

**indicates statistical significance at alpha level of 0.05.*

RESULTS

Descriptive Analysis

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 (2013) software to approximate the average sample size of ∼430 for the French language group for 12 of the booklets. The French groups included all students who received the booklets in French. 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 13 booklets at the $p < 0.01$ level (see **Table 1**).

For twelve booklets, mean total 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 13 booklets, with French language students averaging 3.08 points lower across the 13 booklets. The passage "Fly, Eagle, Fly" is presented in Booklets 1, 2, and 10. The passage "Day Hiking" is presented in Booklets 5, 6, and 10. Reader Booklet 13 includes the passages "Enemy Pie" and "The Giant Tooth Mystery."

Differential Item Functioning Analysis

Prior to conducting the IRT-based DIF detection analyses, fit of the items to IRT models were examined. The Q_1 (Yen, 1993) fit

TABLE 2 | Number of DIF items detected by LH-IRT for booklets 1–13.

		Pro-English			Pro-French			
Booklet	**Total no. of Items**	**Level 2**	**Level 3**	**Total no. for English**	**Level 2**	**Level 3**	**Total no. for French**	**% of DIF items**
1	6	1	4	5	0	1	1	24
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

statistic was used to determine the 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 (Yen and Fitzpatrick, 2006). Since there were no items within the 13 booklets that resulted in poor fit, the Q_1 values are not reported because this indicates that the items fit the IRT models well.

Results of the IRT-DIF detection for all 13 booklets are summarized in **Table 2**. Across the booklets, 11–40% of the items were identified as DIF by the Linn-Harnisch (LH-IRT) based method. Between four and twelve items were identified in total for both languages for each booklet. A total of 50 items were identified as favoring the English language group and 43 as favoring the French language group across the 13 booklets. The overall number of items identified as Level 3 DIF in favor of the English language group was 26 compared to 15 in favor of the French language group across all the booklets.

Results of LR/OLR analysis are shown in **Tables 3** and **4** compare the number of items identified as DIF by the IRT and LR/OLR methods for all the booklets. Some differences were observed in DIF identification between the two methods. The LR/OLR method identified two items, which were not identified as DIF by the LH-IRT method. In addition, even though the LR/OLR DIF method detected DIF for all 74 of the items identified by the IRT method; only 27 of these items were classified as moderate and large DIF. Items flagged as negligible DIF by the LR/OLR methods are reported in the table as showing no DIF. Differences between these two methods were primarily based on the magnitude of the effect size between the two language groups, with the LH method detecting greater effect sizes.

TABLE 3 | Items identified with DIF by LR and OLR methods for booklets 1–13.

Booklet	**Total no. 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
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	6	6	0	Uniform	25
11	1	1	0	Uniform	4
12	1	0	1	Uniform	3
13	4	4	0	Uniform	11

Expert Reviews of Released Items

Bilingual expert reviewers evaluated the equivalence of the French and English language versions of four passages from PIRLS 2011. Passages were reviewed in the sequence of the booklets. In the first stage, passages and items for both language versions were simultaneously reviewed independently, without knowledge about which items were identified statistically as DIF. An equal number of randomly selected non-DIF items were reviewed along with items identified as moderate or large magnitude DIF. In the second stage, the group only reviewed items that were flagged with DIF by both detection methods for which there were rating differences among reviewers. The reviewers were instructed to focus specifically on language, cultural, and format differences and 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. Format and layout differences between language versions that were noted by reviewers are also discussed by passage. For each passage, we provide examples of items given a rating of 2 or 3. **Table 5** contains such examples for the "Fly, Eagle, Fly" passage.

TABLE 4 | Number of DIF items identified by both 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
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	6	0
		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

Three items were statistically flagged with DIF by both the IRT and LR/OLR methods in the "Fly, Eagle, Fly" passage. Expert reviewers evaluated a total of 12 items in this passage, and they identified moderate or large differences for seven items. For three of the four items identified by expert reviewers in this passage, they were unclear which group the differences favored, in part because they identified multiple differences that alternately favored one group or the other. Reviewers evaluated differences between items as being associated with word difficulty and familiarity of vocabulary between the two language tests, wording and sentence structure, and inconsistencies with verb tense. Reviewers also noted that wording and sentence structure differed throughout the text of the passage. Verb tenses differed between language versions, with the English version written in the past tense and the French version written in present tense. Reviewers stated that differences with wording and sentence structure for items in this passage might cause language groups to offer slightly different answers. For example, one of the English items in "Fly, Eagle, Fly" asks examinees to describe what the friend was like while the French version asks to describe his character.

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. The French version was described as confusing and longer than the English version. A total of nine items were reviewed for this passage, with four of those flagged with DIF by both statistical methods. Reviewers identified differences for eight of the nine items they reviewed. The passage "Day Hiking" in English was titled "Discover the fun of day hiking." In the French version it was titled "Découvre les joies de la randonnée", which translates to "Discover the joys of day hiking" in English. Some of the vocabulary differences in this passage were attributed 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. Many of the differences in this passage were attributed 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. Several of the items were worded in the form of a question in the English language version and worded as a statement in the French language version. There was also consensus among the reviewers that key words and phrases varied between the text and questions for the two language versions. 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 "legend" and "tableau qui accompagne la carte," which translates to "chart that accompanies the map" in English. All passage 2 items rated level 2 or 3 by reviewers are displayed in **Table 6**.

Passage 3 was titled "Enemy Pie" in English and "La tarte des ennemis" in French, which translates to "The pie of the enemies" in English. A total of 13 items were reviewed for this passage, and they identified six items as having moderate to large differences. Two items were flagged with DIF by both methods in this passage. Reviewers agreed that in Passage 3 there were many differences with word difficulty and 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," in French, which means "react" in English. 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." Translated to English this sentence reads, "Write one of the ingredients that Thomas expected to find in the pie of the enemies." Overall, with the exception of one item, reviewers' identified the differences as being in the direction of English language students in this passage. As with the first two passages, they attributed many of the differences to poor and inappropriate translation. Differences identified in passage 3 are shown in **Table 7**.

A total of 10 items were reviewed for this passage, and reviewers noted differences with four items. Reviewers identified three items not statistically flagged as DIF as potentially

TABLE 5 | Passage 1 "Fly, Eagle, Fly" expert review ratings and noted differences.

Item	Rating	Favors	Noted differences
		French or English	
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."
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.

problematic. Passage 4 was titled "Giant Tooth Mystery" in English and "Le mystère de la dent Géante" in French. Translated to English the title reads, "The Mystery of the Giant Tooth." There was a general agreement among reviewers that there were differences in words, expressions, and the structure of sentences inherent to Francophone language and culture that made the text in the French version more difficult than the English version, as shown in **Table 8**. For example, the phrase "looked like" was translated to "apparence extérieure" in French, which translates to "external appearance" in English. They also noted differences in word difficulty or familiarity of vocabulary in the French text with phrases such as "caractère intrigant" translated to English as "intriguing character" for the word "puzzling" in the English version. Another example is the use of the phrase "was over 30 m long" in the English version was translated to French as "mesurait 30 m" which translates to English as "measured 30 m." They identified the word "piquant" in the French version as inappropriate translation. The word "piquant" translates to "spicy" in English and was used for the word "spike" in English. Overall, reviewers decided that the differences between the two language versions for this passage were largely due to poor and inappropriate translation.

Correspondence Between DIF Identification and Expert Reviews

Consistency between expert ratings and statistical analysis for all the moderate or large magnitude DIF items that were reviewed are shown in **Table 9**. Items classified as having either moderate or large magnitude differences by reviewers that were not identified as DIF are displayed in the last column.

In the passage, "Fly, Eagle, Fly," expert reviewers identified differences in the two language versions for all three items identified by both methods as DIF. For passage 2 titled, "Day Hiking," expert reviewers again identified all four of the DIF items flagged by both methods. Experts attributed differences for these items to words, expressions, and structure of sentences inherent to the languages; differences in word difficulty or familiarity of vocabulary; differences in meaning; or 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. For passage 3 titled, "Enemy Pie," one of the two items identified as large DIF was identified by expert reviewers. They attributed differences for this item to word difficulty and familiarity of vocabulary. The English version of this item asks, "What does this suggest about the boys?" The French version translated to English asks, "What does this sentence mean to conclude?" In the passage titled "The Giant Tooth," reviewers' rated Item 7 consistent with statistical methods. The use of additional phrases and greater word difficulty attributed to the language differences for Item 7. The English version for Item 7 asks, "What did Gideon Mantell know about reptiles that made the fossil tooth puzzling?" The translated French version asks, "What did Gideon Mantell know about reptiles that makes him understand the intriguing nature of the fossil tooth?" It is important to note, as shown in **Table 6**, that reviewers identified numerous additional items that were either identified by only one of

TABLE 6 | Passage 2 "Day Hiking" expert review ratings and noted differences.

Item	Rating	Favors	Noted differences
		French or English	
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?"
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."
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 "legend" is used but then the phrase "tableau 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.

the statistical methods or by neither as favoring the English language group.

DISCUSSION

International LSAs are 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 (van de Vijver and Leung, 1997; Murat and Rocher, 2004; Tanzer, 2005; Solano-Flores et al., 2006, 2012; Goldstein and Thomas, 2008; Johansson, 2016; Gorur, 2017). 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 and Nelson-Barber, 2001). Measurement differences are not surprising when groups from different countries with different cultures and curricula are compared. Previous research has shown that test adaptation can result in significant score incomparability (Ercikan, 1998, 2003; Gierl and Khaliq, 2001; Maldonado and Geisinger, 2005; Yildirim and Berberoğlu, 2009; Ercikan et al., 2010; Oliveri and von Davier, 2011; Wetzel and Carstensen, 2013; Kreiner and Christensen, 2014). Although research has also demonstrated that psychometric differences between language versions may be attributable to multiple factors (Ercikan and McCreith, 2002; Ercikan et al., 2004a; Sireci et al., 2005; Wu and Ercikan, 2006; Elosua and López-Jaúregui, 2007; Solano-Flores et al., 2009; Arffman, 2010), evidence demonstrates that some differences across groups are attributable to a lack of equivalence across language versions due to translation errors (Oliveri and von Davier, 2011; Ercikan and Lyons-Thomas, 2013; Zhao et al., 2018). This lack of equivalence is, in part, due to test translation procedures (Gierl and Khaliq, 2001; Maldonado and Geisinger,

TABLE 7 | Passage 3 "Enemy Pie" expert review ratings and noted differences.

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

TABLE 8 | Passage 4 "The Giant Tooth Mystery" expert review ratings and noted differences.

Item	Rating	Favors	Noted differences
		English or French	
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 m long, while the French 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."

2005; Yildirim and Berberoğlu, 2009; Arffman, 2010, 2013). The translation of a test does not ensure equivalence between the target and source versions (Beller et al., 2005; Hambleton, 2005; Solano-Flores, 2006; Cohen et al., 2007). In fact, test translation can produce unintended differences in content and difficulty levels between linguistic versions of a test, which may 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, significant differences were reported in the word and character counts between the two versions, and these features were associated with high levels of difficulty for the French versions of the items (Grisay, 2003). Results from the present study are consistent with previous findings that demonstrate differences in difficulty for French and English language versions of LSAs in Canada ranging from 14 to 40% of items (Ercikan, 1998, 2002; Gierl and Khaliq, 2001; Ercikan et al., 2004b, 2010; Oliveri and Ercikan, 2011; Ercikan and Lyons-Thomas, 2013). The use of two DIF detection methods in this study demonstrated that an average of 25% of the items across all 13 PIRLS booklets function differently across language versions, with slightly more items in favor of the English language group than the French language group.

Although some research suggests that it may be impossible to create adapted LSAs that are free from linguistic and cultural bias, specifically for reading literacy tests (Bonnet, 2002; Solano-Flores and Trumbull, 2003; Arffman, 2010), expert review results from this study 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. A majority of the differences between language versions were attributed to poor and

TABLE 9 | 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* (English)	53.70* (uniform)	Moderate (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)	
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)			

(Continued)

TABLE 9 | Continued

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)			

*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 *.*

inappropriate translation. Specific sources of differences identified by reviewers were largely due to length or sentence complexity, word difficulty, familiarity with vocabulary, and differences in meaning.

Findings from this study also align with previous research, indicating that the identification of DIF items can vary with detection methods (Ercikan, 1999; Gierl et al., 1999; Ercikan et al., 2004b; Oliveri and Ercikan, 2011). A number of test characteristics can affect the accuracy of DIF statistics including the range of item difficulties, the distribution of abilities, sample sizes, and differences in procedures for the estimation of DIF and the extent of DIF (Hambleton, 2006; Yildirim and Berberoğlu, 2009). Evidence of inconsistencies between DIF methods supports the use of more than one method to allow for the possibility of simultaneous detection across methods (Hambleton, 2006; Oliveri and Ercikan, 2011; Ercikan and Solano-Flores, 2014). Previous research has also 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 and Lyons-Thomas, 2013). Expert content review can help to identify whether items may be problematic. One of the aims of content review is to allow for the possibility of meaningful interpretations of score differences because the objective of DIF methods is only to statistically flag differences (Puhan and Gierl, 2006; Ercikan et al., 2010). Experts can provide localized linguistic and cultural insight, specific to a country or region with respect to the meaning, relevance, and difficulty of cognitive requirements based on language. Findings from this study demonstrate and support the use of in-country expert reviews to elucidate potential sources of nuanced linguistic and cultural differences.

The results of this study have implications at several phases of testing practices. The first is that test translation procedures for PIRLS in Canada may need to be reexamined to determine if more rigorous and standardized procedures should be adopted. *Standard 9.7* in the *Standards* (American Educational Research Association, 2014) recommends that test developers describe the procedures used to establish and ensure adequacy of translation and adaptation of items and provide evidence of score reliability and validity for linguistic groups. PIRLS International Study Center 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 (Solano-Flores et al., 2009; Arffman, 2010, 2013). Arffman (2013) argues that the process used to translate IEA studies do not necessarily align with the principles and procedures of translation. To address this shortcoming, Arffman recommends that IEA guidelines provide more detailed instructions, with examples to illustrate the delicate tension between two translation purposes that are crucial to international test translation. These two purposes include dynamic translation and equivalence in difficulty. The focus of dynamic translation is to create text that has a similar effect on the target text reader as the source text does on the source text reader. Emphasis is given to the use of natural and authentic language, rather than literal translation. The purpose of difficulty equivalence is to minimize differences across language versions related to required cognitive effort. These dual purposes increase the difficulty of the translation task, and translators are usually not trained to pursue difficulty equivalence (Arffman, 2013). Results from this study and those from Arffman's research on translation procedures utilized by the IEA may indicate that it would be prudent to review current practices for adapting PIRLS in Canada to 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.

The second implication is that test equivalence and score comparability cannot be assumed when tests are adapted for French language groups in Canada (Marotta et al., 2015). Given the increased use of LSAs in Canada, it is imperative that organizations such as the IEA and national testing centers provide evidence of score comparability and evidence of the adequacy and accuracy of actual score interpretations and

uses, to increase the likelihood that inferences, decisions, and consequences are fair, justified and effective (American Educational Research Association, 2014; O'Leary et al., 2017). The IEA and the Council of Ministers of Education (CMEC) must ensure that actual interpretations and uses of scores, as well as actions and consequences based on test scores are justified by evidence to minimize the unintended consequences of legitimate test use (Volante and Jaafar, 2008; O'Leary et al., 2017). Evidence of score comparability and actual interpretations of scores 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 (Labrecque et al., 2012; Klinger et al., 2016). 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, p. 71). Further, the degree of incomparability cannot be generalized across all French language groups in Canada. Test equivalence and score comparability across French and English groups is likely to vary according to the linguistic setting and regional dialect differences within French language groups based on geographical areas, ages, genders, and socio-economic status (Ercikan et al., 2014; Marotta et al., 2015). Some research suggests that linguistic differences between French and English language groups are less extensive for French language students living in a majority setting (Klinger et al., 2016). The linguistic background of students may differentially 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.

The third implication is that results from this study suggest that measurement incomparability between French and English language groups for PIRLS 2011 accounts for some of the observed performance differences in favor of the English language group. At the item level, score comparability was threatened. Expert reviewers found differences between language versions related to words, expressions, meaning, familiarity of vocabulary words, and sentence complexity. Results from this study suggest that caution should be exercised with PIRLS 2011 score comparisons between French and English language groups in Canada.

There were several limitations to this study. The study did not examine diversity within the French and English language samples. The proportion of Canadians that speak either or both of the official languages in Canada varies greatly across provinces. 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 and Lyons-Thomas, 2013). In minority language settings, the composition of French language schools differs substantially. For instance, a number of students attending French-language schools in Ontario immigrated to Canada from African countries such as Somalia, Ethiopia, and Rwanda. 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.

Furthermore, measurement comparability may look different for different subgroups within language categories. 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 and Lyons-Thomas, 2013). Research by Ercikan examines the accuracy of measurement comparability between three French Canadian language groups. The groups include French language students living in majority and minority settings, those living in minority settings that do speak French at home, and those that do not speak French at home. She found 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. Higher reading literacy performance levels were found for Quebec French Francophone students than for French language students living in minority language settings. French language competency was lowest for students attending French language schools living in minority settings who do not speak French at home. Although the current study provides evidence to substantiate measurement incomparability across French and English test versions for LSAs in Canada, it does not address potential differences in measurement comparability within language groups across Canada. Both the number of items and the items identified with DIF are likely to vary for students living in majority and minority settings and for students who speak the test language at home.

Another limitation 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 analyses 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 not statistically flagged with DIF.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://timssandpirls.bc.edu/pirls2011/international-database.html.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the UBC Behavioral Research Ethics Board, certificate number H13-01719. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

This paper was written by SG as part of her thesis requirements. The work was supervised and edited by KE.

REFERENCES

Allalouf, A., Hambleton, R. K., and Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *J. Educ. Meas.* 36, 185–198. doi: 10.1111/j.1745-3984.1999.tb00553.x

Allalouf, A., and Sireci, S. G. (1998). "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, and National Council on Measurement in Education (2014). *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.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scand. J. Educ. Res.* 54, 37–59. doi: 10.1080/00313830903488460

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educ. Meas. Issues Pract.* 32, 2–14. doi: 10.1111/emip.12007

Bachman, L. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Lang. Test.* 17, 1–42. doi: 10.1177/026553220001700101

Baker, C. (1992). *Attitudes & Language.* Philadelphia: Multilingual Matters Ltd.

Beller, M., Gafni, N., and Hanani, P. (2005). "Constructing, adapting, and validating admissions tests in multiple languages: The Israeli Case" in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds R. K. Hambleton, P. F. Merenda, and C. D. Spielberger, Mahwah, NJ: Lawrence Erlbaum Associates.

Bonnet, G. (2002). Reflections in a critical eye: on the pitfalls of international assessment. *Assess. Educ.* 9, 387–399. doi: 10.1080/0969594022000027690a

Boroditsky, L. (2011). How language shapes thought. *Sci. Am.* 304, 62–65. doi: 10.1038/scientificamerican0211-62

Bowles, M., and Stansfield, C. W. (2008). *A Practical Guide to Standards-Based Assessment in the Native Language*. NLA—LEP Partnership.

Campbell, S., and Hale, S. (2003). "Translation and interpreting assessment in the context of educational measurement," in *Translation Today–Trends and Perspectives*, eds G. Anderman and M. Rogers (Toronto: Multilingual Matters Ltd.), 205–224.

Canadian Education Statistics Council (2016). *Education indicators in Canada: An International Perspective 2015*. Available onmline at: https://www150.statcan.gc.ca/n1/en/pub/81-604-x/81-604-x2016001-eng.pdf?st$=$o1VCVhJw

Cohen, A. (2007). "The coming of age for research on test-taking strategies," in *Language Testing Reconsidered,* eds J. Fox, M. Wesche, D. Bayliss, C. Cheng, C. Turner, & C. Doe (Ottawa, ON: University of Ottawa Press), 89–111.

Cohen, Y., Gafni, N., and 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). Available online at: http://www.iaea.info/documents/paper_1162d22ec7.pdf

Crundwell, R. M. (2005). Alternative strategies for large scale student assessment in canada: is value-added assessment one possible answer. *Can. J. Educ. Administr. Policy* 41, 1–21.

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

Cummins, J. (2000). *Language, Power and Pedagogy*: *Bilingual Children in the Cross-Fire.* Clevedon: Multilingual Matters Ltd. doi: 10.21832/9781853596773

Derrida, (1998). *Of Grammatology*. Baltimore, MD. John Hopkins University Press.

Elosua, P., and López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *Int. J. Test.* 7, 39–52. doi: 10.1080/15305050709336857

Ercikan, K. (1998). Translation effects in international assessments. *Int. J. Educ. Res.* 29, 543–553. doi: 10.1016/S0883-0355(98)00047-0

Ercikan, K. (1999). "Translation DIF on TIMSS," in *Annual Meeting of the National Council on Measurement in Education* (Montreal, QC).

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *Int. J. Test.* 2, 199–215.

Ercikan, K. (2003). Are the English and French versions of the third international mathematics and science study administered in Canada comparable? Effects of adaptations. *Int. J. Educ. Policy Res. Pract.* 4, 55–76.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., and Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educ. Meas.* 29, 24–35. doi: 10.1111/j.1745-3992.2010.00173.x

Ercikan, K., Domene, J. F., Law, D., Arim, R., Gagnon, F., and 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., and Koh, K. (2004b). Comparability of bilingual versions of assessments: sources of incomparability of English and French versions of Canada's national achievement tests. *Appl. Meas. Educ.* 17, 301–321. doi: 10.1207/s15324818ame1703_4

Ercikan, K., and Lyons-Thomas, J. (2013). "Adapting tests for use in other languages and cultures," in *APA Handbooks in Psychology. APA Handbook of Testing and Assessment in Psychology, Vol. 3. Testing and Assessment in School Psychology and Education,* eds K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez, 545–569. doi: 10.1037/14049-026

Ercikan, K., and McCreith, T. (2002). Disentangling sources of differential item functioning in multi-language assessements. *Int. J. Test.* 2, 199–215. doi: 10.1207/S15327574IJT023&4_2

Ercikan, K., Roth, W. M., Simon, M., Sandilands, D., and Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Appl. Meas. Educ.* 27, 273–285. doi: 10.1080/08957347.2014.944306

Ercikan, K., and Solano-Flores, G. (2014). Introduction to the special issue: levels of analysis in the assessment of linguistic minority students. *Appl. Meas. Educ.* 27, 233–235. doi: 10.1080/08957347.2014.944462

Fairbairn, S. B., and Fox, J. (2009). Inclusive achievement testing for linguistically and culturally diverse test takers: essential considerations for test developers and decision makers. *Educ. Meas.* 28, 10–24. doi: 10.1111/j.1745-3992.2009.01133.x

Gee, J. P. (2001). Reading as situated language: A sociocognitive perspective. *J. Adolesc. Adult Literacy*, 44, 714–725. doi: 10.1598/JAAL.44.8.3

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Can. J. Educ.* 25, 280–96. doi: 10.2307/1585851

Gierl, M. J., and Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *J. Educ. Meas. Summer* 38,164–187. doi: 10.1111/j.1745-3984.2001.tb01121.x

Gierl, M. J., Rogers, W. T., and Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta J. Educ. Res.* 45:353.

Goldstein, H., and Thomas, S. M. (2008). Reflections on the international comparative surveys debate. *Assess. Educ.* 15, 215–222. doi: 10.1080/09695940802417368

Gorur, R. (2017). Towards productive critique of large-scale comparisons in education. *Critic. Stud. Educ.* 58, 341–355. doi: 10.1080/17508487.2017.1327876

Greenfield, P. M. (1997). You can't take it with you: why ability assessments don't cross cultures. *Am. Psychol.* 52, 1115–1124. doi: 10.1037/0003-066X.52.10.1115

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Lang. Test.* 20, 225–240. doi: 10.1191/0265532203lt254oa

Grisay, A., and Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Stud. Educ. Eval.* 33, 69–86. doi: 10.1016/j.stueduc.2007.01.006

Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguist. Educ.* 5, 93–116. doi: 10.1016/0898-5898(93)90026-7

Hambleton, R. K. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* eds R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum Associates), 3–38. doi: 10.4324/9781410611758

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Med. Care* 44, S182–S188.

He, J., and van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Psychol. Cult.* 2:1111. doi: 10.9707/2307-0919.1111

International Association for the Evaluation of Educational Achievement (2012). *International Database Analyzer (version 3.0)*. Hamburg: IEA Data Processing and Research Center.

Johansson, S. (2016). International large-scale assessments: what uses, what consequences?. *Educ. Res.* 58, 139–148. doi: 10.1080/00131881.2016.1165559

Klinger, D., DeLuca, C., and Merchant, S. (2016). "Canada: The intersection of international achievement testing and educational policy development." *The Inter-section of International Achievement Testing and Educational Policy: Global Perspectives on Large-Scale Reform,* ed L. Volante (New York, NY: Routledge Press), 140–159.

Kreiner, S., and Christensen, K. B. (2014). Analyses of model fit and robustness: a new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika* 79, 210–231. doi: 10.1007/s11336-013-9347-z

Labrecque, M., Chuy, M., and Brochu, P. (2012). *Canada in Context: Canadian Results from the Progress in International Reading Literacy Study. 9780889872233.* Toronto, ON: Council of Ministers of Education, Canada, 2012.

Linn, R. L., and Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *J. Educ. Meas.* 18, 109–118. doi: 10.1111/j.1745-3984.1981.tb00846.x

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ, Lawrence Erlbaum Associates.

Maldonado, C. Y., and Geisinger, K. F. (2005). "Conversion of the wechsler adult intelligence scale into Spanish: An early test adaptation effort of considerable consequence," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* eds R. K. Hambleton, P. F. Merenda and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum Associates), 213–234.

Marotta, L., Tramonte, L., and Willms, J. D. (2015). Equivalence of testing instruments in Canada: studying item bias in a cross-cultural assessment for preschoolers. *Can. J. Educ.* 38, 1–23.

Miller, T. R., and Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *J. Educ. Meas.* 30, 107–122. doi: 10.1111/j.1745-3984.1993.tb01069.x

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

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *ETS Res. Rep. Ser.* 1992, i–30. doi: 10.1177/014662169201600206

Murat, F., and Rocher, T. (2004). "On the methods used for international assessments of educational competences," in *Comparing Learning Outcomes: International Assessment and Education Policy,* eds J. H. Moskowitz and M. Stephens (London: Routledge Falmer), 190–214.

O'Leary, T. M., Hattie, J. A., and Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educ. Meas.* 36, 16–23. doi: 10.1111/emip.12141

Oliveri, M. E., and Ercikan, K. (2011). Do different approaches to examining construct comparability lead to similar conclusions? *Appl. Meas. Educ.* 24, 1–18. doi: 10.1080/08957347.2011.607063

Oliveri, M. E., Olson, B., Ercikan, K., and Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *Int. J. Test.* 12, 203–223. doi: 10.1080/15305058.2011.617475

Oliveri, M. E., and von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychol. Test Assessm. Model.* 53:315.

Puhan, G., and Gierl, M. J. (2006). Evaluating the effectiveness of two-stage testing on English and French versions of a science achievement test. *J. Cross-Cult. Psychol.* 37:136. doi: 10.1177/0022022105284492

Rogers, W. T., Lin, J., and Rinaldi, C. M. (2010). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Appl. Meas. Educ.* 24, 39–70. doi: 10.1080/08957347.2011.532416

Rorty, R. (1977). Derrida on language, being and abnormal philosophy. *J. Philos.* 74, 673–681. doi: 10.2307/2025769

Roth, W. M. (2009). Realizing Vygotsky's program concerning language and thought: tracking knowing (ideas, conceptions, beliefs) in real time. *Lang. Educ.* 23, 295–311. doi: 10.1080/09500780902954240

Roth, W. M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., and Ercikan, K. (2013).Investigating linguistic sources of differential item functioning using expert think-aloud protocols in science achievement tests. *Int. J. Sci. Educ.* 35, 546–576. doi: 10.1080/09500693.2012.721572

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

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

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teach. College Record* 108, 2354–2379. doi: 10.1111/j.1467-9620.2006.00785.x

Solano-Flores, G., Backhoff, E., and Contreras-Niño, L. Á. (2009). Theory of test translation error. *Int. J. Test.* 9, 78–91. doi: 10.1080/15305050902880835

Solano-Flores, G., Contreras-Niño, L., and Backhoff, E. (2006). Translation and adaptation of tests: lessons learned and recommendations for countries participating in TIMSS, PISA and other international comparisons. *Rev. Electr. Investig. Educ.* 8:2.

Solano-Flores, G., Contreras-Niño, L. Á., and Backhoff, E. (2012). "The measurement of translation error in PISA-2006 items: an application of the theory of test translation error," in *Research on PISA*, eds M. Prenzel, M. Kobarg, K. Schops, and S. Ronnebeck (Dordrecht: Springer), 71–85. doi: 10.1007/978-94-007-4458-5_5

Solano-Flores, G., and Nelson-Barber, S. (2000). "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., and Nelson-Barber, S. (2001). On the cultural validity of science assessments. *J. Res. Sci. Teach.* 38, 553–573. doi: 10.1002/tea.1018

Solano-Flores, G., and Trumbull, E. (2003). Examining language in context: the need for new research and practices paradigms in the testing of English-language learners. *Educ. Res.* 32, 3–13. doi: 10.3102/0013189X032002003

Steiner, V., and Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educ. Psychol.* 31, 191–206. doi: 10.1207/s15326985ep3103&4_4

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *J. Educ. Meas.* 27, 361–370. doi: 10.1111/j.1745-3984.1990.tb00754.x

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

Tobin, M., Nugroho, D., and Lietz, P. (2016). Large-scale assessments of students' learning and education policy: synthesising evidence across world regions. *Res. Papers Educ.* 31, 578–594. doi: 10.1080/02671522.2016.1225353

van de Vijver, F. J., and Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks, CA: SAGE Publications, Incorporated.

Volante, L., and Jaafar, S. B. (2008). Educational assessment in Canada. *Assess. Educ.* 15, 201–210. doi: 10.1080/09695940802164226

Wetzel, E., and Carstensen, C. H. (2013). Linking PISA 2000 and PISA 2009: implications of instrument design on measurement invariance. *Psychol. Test Assess. Model.* 55, 181–206.

Wu, A., and Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International J. Test.* 6, 287–300. doi: 10.1207/s15327574ijt0603_5

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *J. Educ. Meas.* 30, 187–214. doi: 10.1111/j.1745-3984.1993.tb00423.x

Yen, W. M., and Fitzpatrick, A. R. (2006). Item response theory. *Educ. Meas.* 4, 111–153.

Yildirim, H. H., and Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *Int. J. Test.* 9, 108–121. doi: 10.1080/15305050902880736

Zhao, X., Solano-Flores, G., and Qian, M. (2018). International test comparisons: eviewing translation error in different source language-target language combinations. *Int. Multiling. Res. J.* 12, 17–27. doi: 10.1080/19313152.2017.1349527

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: Directorate of Human Resources Research and Evaluation, Department of National Defense. Available online at: http://educ.ubc.ca/faculty/zumbo/DIF/index.html

Zumbo, B. D. (2008). *Statistical Methods for Investigating Item Bias in Self-Report Measures*. Florence: Universita' degli Studi di Firenze E-prints Archive. Available online at: http://eprints.unifi.it/archive/00001639

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.