APPLIED MEASUREMENT IN EDUCATION, *17*(3), 301–321

Comparability of Bilingual Versions of Assessments: Sources of Incomparability of English and French Versions of Canada's National Achievement Tests

Kadriye Ercikan
Department of Educational & Counseling Psychology and Special Education
University of British Columbia

Mark J. Gierl
Department of Educational Psychology
University of Alberta

Tanya McCreith
Department of Educational & Counseling Psychology and Special Education
University of British Columbia

Gautam Puhan
Department of Educational Testing Service
University of Alberta

Kim Koh
Department of Educational & Counseling Psychology and Special Education
University of British Columbia

This research examined the degree of comparability and sources of incomparability of English and French versions of reading, mathematics, and science tests that were

Requests for reprints should be sent to Kadriye Ercikan, University of British Columbia, Educational and Counseling Psychology and Special Education, Faculty of Education, 2125 Main Mall, B.C. Canada V6T 1Z4. E-mail: kadriye.ercikan@ubc.ca

administered as part of a survey of achievement in Canada. The results point to substantial psychometric differences between the 2 language versions. Approximately 18% to 36% of the items were identified as differentially functioning for the 2 language groups. Large proportions of these differential item functioning (DIF) items, 36% to 100% across age groups and content areas, were attributed to adaptation related differences. A smaller proportion, 27% to 33% of the DIF items, was attributed to curricular differences. Twenty-four to 49% of DIF items could not be attributed to either of the 2 sources considered in the study.

During the last decade, test adaptations and translations have become prevalent because of an increase in international testing, more demand for credentialing and licensure exams in multiple languages, and a growing concern to test students in their first language. For example, the International Association for the Evaluation of Educational Achievement conducted the Third International Mathematics and Science Study (TIMSS) in 1995 and 1999 by administering tests in more than 40 different languages. The Council of Ministers of Education in Canada assesses the achievement of 13- and 16-year-old students in reading and writing, mathematics, and science in English and French for the provinces and territories as part of the School Achievement Indicators Program (SAIP). In Israel, university entrance examinations are adapted from Hebrew to five other languages. In the United States, New York Regents Examinations of Mathematics are offered in five languages in addition to English. The National Assessment of Educational Progress was administered in Spanish in addition to English for the first time in 2003. The comparability of test results across different language versions of these tests is at the core of the validity of interpretations in these assessments. This comparability is the focus of this study.

Efforts to create tests that are as similar as possible across different languages involves not only a translation that preserves the original test meaning but additional changes such as those affecting item format and testing procedures may be necessary to insure equivalence of the versions of the test in multiple languages (Hambleton, 1993). In achievement tests in particular, expectations regarding familiarity of examinees with different measurement units, such as inches versus centimetres, and notations, such as13:00 hr versus 1 p.m., are considered when tests are translated. This more general process of converting one language version of a test to another is defined as test adaptation. Therefore, in this article, we chose to use the term *adaptation* instead of *translation* to communicate the process involved in converting tests from one language to another more accurately.

The assumption that multiple language test forms, even when developed by a group of testing specialists and bilingual experts, will measure comparable constructs is questionable without an empirical verification of such comparability. A poor adaptation can affect the meaning of test items and adversely influence the comparability and interpretability of test scores across language groups.

Hambleton (1994) provided one illustrative example. In a Swedish–English comparison, English-speaking examinees were presented with this item:

> Where is a bird with webbed feet most likely to live?
> (a) in the mountains
> (b) in the woods
> (c) in the sea
> (d) in the desert. (p. 235)

In the Swedish adaptation the phrase "webbed feet" became "swimming feet" thereby providing an obvious clue to the Swedish-speaking examinees about the correct option for this item. Therefore, what is assessed by the item and the difficulty level of the item is altered by adaptation to Swedish. Previous research on multilingual examinations has demonstrated that test adaptation can affect comparability, and therefore, validity, and fairness for groups taking the test in different languages (e.g., Angoff & Cook, 1988; Sireci & Berberoglu, 2000; Sireci, Fitzgerald, & Xing, 1998; van der Vijver & Tanzer, 1998). Recent research conducted by Ercikan (1998, 1999), Ercikan and McCreith (2002), Gierl, Rogers, and Klinger (1999), and Gierl and Khaliq (2001) using tests taken by English- and French-speaking Canadian students found that differences due to adaptations were associated with psychometric differences between two language versions of tests as well. These researchers also demonstrated that psychometric differences between the language versions of tests may not necessarily be due to adaptation differences, and other factors may affect item equivalence across language versions of tests. These, for example, include cultural and curriculum differences between the groups. Cultural differences can influence examinees' intrinsic interest in and familiarity with the content or context of items. Previous research on item equivalence for gender and ethnic groups has shown that the context used in an item can affect this equivalence (O'Neil & McPeek, 1993). Curriculum-related differences can result in varying degrees of student exposure to the domain of items depending on the student's country. The order in which topics are introduced (e.g., Algebra, Fractions, and Number Sense) and subsequent instruction provided to students could be different in comparison groups. For example, if one of the groups has not covered Algebra when the assessment took place, then this group would be expected to do more poorly relative to their performance on the rest of the test. In other words, certain topics or subjects will undoubtedly have differential coverage in different countries, and differential coverage can lead to differential response patterns and difficulty levels, independent of any problems due to adaptation.

TEST ADAPTATION AND DIFFERENTIAL ITEM FUNCTIONING

The comparability of test items across different groups is often evaluated using differential item functioning (DIF) analyses. DIF is present when examinees from different groups have a different probability or likelihood of answering an item correctly, after conditioning on overall ability. DIF analyses can be conducted during or after the test adaptation process to identify items that function differently between language groups. Previous research highlights three aspects of such analyses that need to be considered in designing studies for examining comparability of test results across language groups. First, the amount of DIF on some adapted tests is large. For example, Ercikan (1999) found that 58 out of 140 science items (41%) and 29 out of 158 mathematics items (18%) from TIMSS displayed DIF when the Canadian English and French examinees were compared. Similarly, Gierl et al. (1999) reported that 26 of 50 items (52%) on a Canadian Grade 6 social studies achievement test adapted from English to French displayed DIF. These findings highlight that comparability is not ensured by simply adapting tests into multiple languages; rather, the language forms must be analyzed using DIF methods to ensure they yield comparable results.

Second, previous research has demonstrated that the incomparability of constructs identified in multilingual assessments is not necessarily due to differences created by the adaptations. For example, Ercikan and McCreith (2002), in their examination of the comparability of English and French versions of the TIMSS test items administered in Canada, found that DIF could be attributed to adaptation related differences in approximately 22% of the mathematics DIF items and 40% of the science DIF items. These findings point to the importance of examining the sources of DIF.

The third finding in relation to DIF analyses that needs to be considered in examining comparability of test results across language groups is that identification of items as DIF can vary depending on which DIF detection method is used. For example, Gierl et al. (1999) used the Mantel–Haenzsel (M–H) and Simultaneous Item Bias Test (SIBTEST) DIF detection procedures to identify items on a social studies test that may be functioning differentially for English- and French-speaking examinees. M–H identified 19 DIF items, whereas SIBTEST identified 27 DIF items on the 55-item test. Moreover, all 19 items identified by M–H were also identified by SIBTEST suggesting that M–H is a more conservative DIF detection method. Similarly, Ercikan (1999) in examining comparability of English and French versions of TIMSS items administered in Canada using the logistic regression and item response theory (IRT) based DIF detection methods, found moderate consistency between the two DIF detection procedures. The logistic regression based DIF detection procedure identified 19% more DIF items than the IRT based DIF detection procedure. These results strongly suggest that at least

two DIF procedures should be used to establish a consistent and defensible pattern of DIF when attempting to identify items that function differentially between language groups. These three findings guided the research design used in this study.

IDENTIFYING SOURCES OF DIF

Even though DIF identification methods are commonly used for examining construct comparability of tests for different groups, one of the limitations of these methods is related to identification of sources of DIF. Ercikan (2002) reviewed success levels in identifying sources of DIF and reported mixed levels of success in these studies. She identified three factors that might affect success levels in identifying sources of DIF using content or bilingual experts. The first is whether the reviewers have knowledge about DIF status of items. When the reviewers have knowledge about the DIF status of items, the item review process becomes identifying item characteristics that might be related to differences on the DIF items only. This process does not allow any way of checking for accuracy of interpretations and may lead to inflated success rates.

The second factor is whether a single or multiple versions of tests are reviewed. In multilanguage versions of tests, reviews of comparability of tests focuses on comparability of format, content, and language and targets to identify more obvious differences. Reviews of a single version of DIF items for gender and ethnic groups focus on context and content that might make the item biased for a group. This type of a review attempts to identify subtleties in item content and context that might lead to differential cognitive processes among comparison groups and is expected to be more complex.

The third factor that might affect success rates in identifying sources of DIF is the number of potential sources of DIF. Different numbers of sources of DIF might be at play even in multiple language versions of assessments depending on whether a test is an achievement test, a licensure test, and whether the test is administered in one country or multiple countries. For example, in achievement tests, performance on the test is expected to be closely tied to the curricular coverage of the item topic. Therefore, differences in curriculum and instruction are expected to be important factors affecting DIF in achievement tests. Licensure tests, on the other hand, are designed to assess minimum competency and performance is expected to be affected by curricular differences to a lesser degree. Therefore, curricular differences in licensure tests may not be important sources of DIF. The larger numbers of potential sources of DIF are expected to make the identification and disentangling of sources of DIF more complex.

The implication of DIF and differences in constructs assessed for the groups cannot be examined without knowledge about the sources of DIF. Therefore, one of the important components of this study is identification of sources of DIF.

PURPOSE OF THIS STUDY

The focus of this article is the comparability of multiple language versions of tests and in particular English and French versions of Canadian SAIP tests. The comparability issues examined in this study exist in international assessments or other multilingual versions of assessments of achievement. In addition to providing results regarding the degree of comparability that can be expected in such assessments, this study demonstrates methodology used in examining construct comparability and identifying sources of incomparability in multiple language versions of tests. Two sources of differences, adaptation related and curricular, in constructs assessed by the two language versions of tests are considered. These two sources are expected to be important differences between the two language groups that may lead to DIF. This article describes and discusses strategies used in examining comparability of multiple language versions of tests and identifying sources of DIF in multilanguage assessments in which multiple factors are expected to be causing DIF. Three strategies are used for identifying adaptation and curricular differences as sources of DIF: (a) judgmental reviews by multiple bilingual translators of all items, (b) cross-validation of DIF in two age groups, and (c) examination of the distribution of DIF items by curricular topic area. Other major factors that might affect DIF, such as cultural and instructional differences, are not examined, therefore interpretations of DIF are not exhaustive. Rather, an evidence gathering approach is taken where analyses are conducted to gather evidence that may provide support for or against the following hypotheses:

1. DIF is due to adaptation effects: Two types of evidence were used to support this hypothesis. First was identification of differences in meaning, structure, and format between translated versions of items in judgmental reviews. Second was the cross-validation of DIF in one other comparison.
2. DIF is due to curricular differences: Clustering of DIF items that are all in favor or against the same group in a curricular topic area, such as Algebra, was interpreted as evidence supporting the hypothesis that there was an association between DIF and that topic area. In other words, DIF may be due to curricular differences in relation to this topic.

METHOD

Data

Data from Canada's national examination, the SAIP, developed by the Council of Ministers of Education (CMEC), were used for evaluating the comparability of test items and assessment results from the English and French versions of tests in

three content areas. SAIP assesses 13-and 16-year old students in the areas of language arts, mathematics, and science. The examinations in this testing program are developed in both English and French and are designed to be equivalent for the English- and French-speaking examinees. According to the information presented in the SAIP reports about the bilingual test development process, items were developed by both English- and French-speaking test developers using a simultaneous test development model given in the following description:

> From the outset, the instruments were developed by English- as well as by French-speaking educators working together for the purpose of eliminating any possible linguistic bias. Whether they wrote in French or in English, the students responded to the same questions and executed the same tasks. Consequently, the statistical results presented for each language group in this report can be compared with reasonable confidence (CMEC , 2000).

No empirical analyses are presented to support the assumption that the English and French forms are parallel.

This study presents results from the mathematics, reading, and the science assessments, administered in 1997, 1998, and 1999, respectively. Each assessment is administered to a random sample of 13-year-olds and 16-year-olds from all provinces and territories. Each of the assessments is briefly described next.

Reading. The SAIP Report on Reading and Writing Assessment (CMEC,1998) indicated that this assessment included a booklet of readings from recognized literature, essays, and newspaper articles. Some readings are complete articles and some are excerpts of longer works. The selections represented varying lengths (up to four pages), different genres, and various degrees of difficulty. After reading the materials, students were asked to answer multiple-choice questions and also to respond in writing to specific questions and tasks. During the reading assessment, students are presented with three types of questions: (a) interpretive questions that require students to demonstrate an understanding of the reading passages at literal and figurative levels, (b) evaluative questions that ask students to make judgments about textual information and the author's purposes, and (c) extension and extrapolation questions that require the student to relate concepts in the texts to their personal experiences, explaining the links clearly. There were three forms, A, B, and C, of the SAIP Reading tests, that were randomly assigned to the examinees. The two language versions of Forms B and C had different passages and associated questions in the two languages and therefore were not comparable. As a result, only Form A, which had 22 multiple-choice items, was used in examining the comparability of the English and French versions of Reading tests in this study.

Mathematics. The mathematics assessment was designed to evaluate students' understanding of mathematics content including their knowledge of numbers and operations, algebra and functions, measurement and geometry, and data management and statistics. The SAIP Report on Mathematics Assessment (CMEC, 1997) indicated that there were two test booklets: Booklet 1 contained 27 background questions, 15 multiple-choice placement questions, and 110 questions grouped in five sections according to levels of performance; Booklet 2 contained space for recording answers. All students began with the background questions and then moved on to the placement test. When the 15-question placement test was completed, students raised their hands to indicate this, and the supervising teacher, using a template over the appropriate section, immediately scored their responses. Students who scored 0 to 10 were to begin the 110 questions at Question 16 (Section A). Students who scored 11 to 13 were to begin at Question 41 (Section B). Students who scored 14 or 15 were to begin at Question 66 (Section C). The 110 questions were a combination of multiple-choice and constructed-response questions.

Science. The science assessment was designed to test general knowledge and concept of science and inquiry skills, as well as the relation of science to technology and societal issues. The SAIP Report on Science Assessment (CMEC, 1999) indicated that the written component of the assessment included multiple-choice and constructed-response questions related to general scientific knowledge. All students writing this assessment began by doing 12 questions, which constituted Form A. On the basis of their scores on those 12 questions, students were directed to a subsequent particular set of color-coded pages in their test booklet. Those examinees who scored less than 8 were assigned to Form B and those who scored 8 or more were assigned to Form C. Each form contained 66 items that were a combination of multiple-choice and constructed-response questions. Examinees were only asked to complete the 66 items in their assigned booklet (unlike math, in which the examinees were asked to write as many items as they could in the allotted time). The sample size and numbers of items by item type for each assessment are presented in Table 1.

Identification of DIF

Two DIF detection methods were used to identify items on the SAIP reading, mathematics, and science achievement tests that may be functioning differentially between English- and French-speaking examinees. DIF items were identified using an application of the Linn–Harnisch (L–H) method (Linn & Harnisch, 1981) to IRT-based item parameters and the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993). These two approaches are not expected to give identical results but are used to verify and confirm the DIF status of the items analyzed.

TABLE 1
Sample Size and Number of Items for the SAIP Reading, Mathematics, and Science Assessments

				Number of Students			
	Number of Items			*13-year-olds*		*16-year-olds*	
Content Area	*MC*	*CR*	*Total*	*English*	*French*	*English*	*French*
Reading							
Form A	22	0	22	3,230	1,097	2,934	959
Mathematics	75	50	125	9,029	3,509	8,104	2,719
Science							
Form A	10	2	12	8,961	3,166	8,263	3,024
Form B	40	26	66	4,585	1,581	2,296	904
Form C	40	26	66	4,362	1,548	5,924	2,114

Note. SAIP = School Achievement Indicators Program; MC = multiple choice; CR = constructed response.

Due to the multistage nature of SAIP, in some of the test forms, only portions of items are administered to students. This administration procedure creates missing data for these forms that cannot be analyzed using SIBTEST. In the IRT based DIF procedure, however, the responses to items that are not administered to examinees can be considered as missing (not-reached items) and included in the analyses. Consequently, items on the mathematics test (which draws heavily on the multistage testing approach) were analyzed only with L–H. The two methods are described next.

IRT-based L–H. IRT provides a coherent conceptual framework for studying DIF (Ackerman, 1992; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). Lord (1952) initially proposed a theory based on the item characteristic curve (ICC) that allowed researchers to model the relation between an unobservable latent trait believed to characterize test performance and the probability that an examinee would correctly solve a test item. ICCs facilitate the comparison of item level performance across groups after controlling for differences in group ability. If a test item has the same ICC for every group, then examinees of the same ability level will have the same chance of correctly answering the item. Alternatively, if a test item has a different ICC for one group compared to another, then the item is functioning differently across the groups (Lord, 1980). Several statistical procedures are used to quantify and test for the significance of the differences between group ICCs. One of the simplest procedures for quantifying DIF and testing for significance is the L–H statistic (Linn & Harnisch, 1981). This procedure is used to compute, for each item, the observed and expected mean response and the

difference between them (observed minus predicted) by deciles of the specified subgroup and for the subgroup as a whole. The expected values are computed using the parameter estimates obtained from the entire sample, and the θ estimates (ability estimates) for the members of the specified subgroup. The differences between observed and expected mean responses are used to calculate a chi-square statistic. For large sample sizes (greater than 30), the chi-square statistics with k degrees of freedom can be approximated by the Standard Normal Distribution using $Z_p = (\chi^2{}_p - k) / (2k)^{1/2}$ where Z_p is the pth percentile of the standard normal distribution. The items with Z-statistic greater than 2.58 were identified as functioning significantly differently for the two comparison groups at $\alpha = 0.005$ level. Different degrees of DIF classified as follows: An item is classified as Level 3 DIF if the absolute value of the obtained minus the expected mean is greater than 0.10, and also the corresponding absolute Z value is $|Z| > 2.58$. If $|Z| > 2.58$ but the expected mean difference is less than 0.10 then the item is classified as a Level 2 DIF item. Items with $|Z| < 2.58$ are classified as Level 1, which indicates that the item is DIF free.

In the L–H procedure, multiple-choice items were calibrated using the three-parameter logistic (3-PL) model (Lord, 1980) and the open-ended items were calibrated using the two-parameter partial credit (2-PPC) model (Yen, 1993). The 2-PPC model is a special case of Bock's (1972) nominal model and is equivalent to Muraki's (1992) generalized partial credit model. Similar to the generalized partial credit model, in 2-PPC, items can vary in their discriminations and each item has location parameters, one less than the number of score levels. The calibrations were conducted simultaneously for the two item types using marginal maximum likelihood estimation procedure. This procedure was implemented using PARDUX, an IRT calibration and analysis software developed by CTB/McGraw-Hill (1991).

SIBTEST. DIF statistical analyses were conducted for the reading and science test items using the SIBTEST. SIBTEST is a nonparametric method for detecting differential item and test functioning that was developed as an extension of Shealy and Stout's (1993) multidimensional model for DIF. In this framework, DIF is conceptualized as a difference in the probability of selecting a correct response, which occurs when individuals in groups with the same levels of the latent attribute of interest (θ), possess different amounts of nuisance abilities (h) that might influence their item response patterns.

The statistical hypothesis tested by SIBTEST is:

$$H_0: B(T) = P_R(T) - P_F(T) = 0$$
$$vs.$$
$$H_1: B(T) = P_R(T) - P_F(T) \neq 0,$$

where $B(T)$ is the difference in probability of a correct response on the studied item for examinees in the reference and focal groups matched on true score; $P_R(T)$ is the probability of a correct response on the studied item for examinees in the Reference group with true score T; and $P_F(T)$ is the probability of a correct response on the studied item for examinees in the focal group with true score T . With the SIBTEST procedure, items on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest contains items that are suspected of having DIF and the matching subtest contains items that, ideally, are known to be unbiased and measure only the primary dimension on the test. Linear regression is used to estimate corresponding subtest true score for each matching subtest score. These estimated true scores are adjusted using a regression correction technique to ensure the estimated true score is comparable for the examinees in the reference and focal groups on the matching subtest (Shealy & Stout, 1993). In the final step, $B(T)$ is estimated using $\hat{B}$, which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels. SIBTEST provides an overall statistical test and a measure of the effect size for each item ($\hat{B}$ is an estimate of the amount of DIF). According to Roussos and Stout (1996, p. 220) the following $\hat{B}$ values are used for classifying DIF as negligible, moderate, and large:

- Negligible or Level 1 DIF: Null hypothesis is rejected and $|\hat{B}| < 0.059$.
- Moderate or Level 2 DIF: Null hypothesis is rejected and $0.059 \leq |\hat{B}| < 0.088$.
- Large or Level 3 DIF: Null hypothesis is rejected and $|\hat{B}| \geq 0.088$.

Adaptation Review Process

Four bilingual French–English translators completed a blind review of the SAIP items for identifying potential sources of DIF and adaptation problems. The translators were fluent in both languages and had extensive experience in teaching. Their teaching experiences ranged between 5 to 10 years at high school levels teaching in the areas of mathematics, science, English-as-a-second-language and French-as-a-second-language. The adaptation review process not only requires the identification of differences in the two language versions but judgments regarding whether the differences are expected to lead to performance differences for the two language groups as well. Therefore, experience in teaching and familiarity with student thinking processes were considered to be important characteristics of translators as well. The translators were asked to evaluate the equivalence of English and French versions of test items and rate their equivalence according to the following criteria:

0—No difference in meaning between the two versions;
1—Minimal differences in meaning between the two versions;

2—Clear differences in meaning between the two versions but they may not necessarily lead to differences in performance between two groups;
3—Clear differences in meaning between the two versions that are expected to lead to differences in performance between two groups.

The review process consisted of three stages: (a) group review of sample of items to discuss and understand criteria involved in reviewing the items, (b) independent review of each item by four reviewers, and (c) group discussion and consensus for rating adaptation differences between the two language versions of the items. Form A of the reading items, a random sample of mathematics items, which included all mathematics DIF items, and all of the science items were reviewed.

RESULTS

The comparability of the English and French versions of SAIP test items were conducted separately for 13- and 16- year-olds, for each of the content areas, reading, mathematics, and science. These analyses used L–H and SIBTEST methods to identify differentially functioning test items for the two language groups. The replication of DIF analyses for the two age groups contributed to our evaluation of the comparability of the two language versions of SAIP in two ways: (a) the verification of findings from one age group with another and (b) the relative comparability of English and French versions of SAIP for the two age groups. DIF captures similarity of constructs being assessed in the two comparison groups. This similarity is expected to be somewhat different in each of the age group comparisons due to differing degrees of similarities with curriculum and instruction. In addition, small differences in adapted versions of tests may lead to DIF in one age group but not the other due to differences in ability distributions as well as differential effects of item language, content, and format on examinee performance in that population. However, if DIF is replicated in both age group comparisons, it can provide an additional source of evidence for supporting the interpretation that DIF may be due to adaptation differences.

The following sections summarize findings regarding the comparability of DIF results using the two DIF detection methods, the degree of DIF in the English and French versions of the items, and evidence supporting whether DIF was due to adaptation effects or curricular effects.

Comparability of L–H and SIBTEST Results

The two DIF detection methods were used to identify items in Form A of reading and across all forms in science. Only the L–H method was used to identify DIF in

TABLE 2
Classification of DIF by L–H and SIBTEST

		Reading[a]				*Science*[c]		
		SIBTEST DIF Level				*SIBTEST DIF Level*		
Age Group	*L–H DIF Level*	*1*	*2*	*3*	*Mathematics*[b]	*1*	*2*	*3*
13-year-olds	1	11	1	2	78	63	0	2
	2	4	2	2	45	27	15	27
	3	0	0	0	2	0	0	10
16-year-olds	1	11	1	0	85	73	0	2
	2	3	4	3	37	20	19	20
	3	0	0	0	3	0	0	10

Note. DIF = differential item functioning; L–H = Linn–Harnisch; SIBTEST = Simultaneous Item Bias Test.
[a]22 items. [b]125 items. [c]144 items.

Mathematics because the multistage administration design for this test created a sparse data matrix, that could not be analyzed using SIBTEST. Overall, the L–H method consistently identified more items compared to SIBTEST. This difference in the number of flagged items ranged from 5% to 17% of the number of possible items, depending on content area, and the age of the students. In reading, there was only a difference of one item, for both ages. Conversely, in the science assessment for the 13-year-olds, the L–H method flagged 25 more items than SIBTEST. All items that were flagged by both detection methods were consistent in their indication of which population, the English- or the French-speaking students, were favored by a particular DIF item, although the degree of DIF, whether a Level 2 or 3 DIF, was often different. Specifically, when there was a difference in the degree of DIF, SIBTEST consistently would indicate a higher level of DIF. Table 2 provides a summary of the number of items identified by each statistical procedure, and the degree of overlap, separately by content area and age.

Reading (Form A). A large percentage of reading items were identified as DIF. Using the L–H method, 8 of 22 items (36%) were identified as DIF for 13-year-olds, and 10 of 22 items (45%) were identified as DIF for the 16-year-olds. All of these DIF items were at Level 2. At least half of the DIF items, 63% for 13-year-olds and 50% for 16-year-olds, favored the French-speaking students. The number of DIF items that were identified as DIF by SIBTEST was one less than that from the L–H analyses for each age group. For the 13-year-olds, 7 items (32%) were identified as DIF, 4 of which were at Level 3. For the 16-year-olds, 9 items (41%) were identified as DIF, 3 of which were at Level 3. Fewer than half of these

DIF items, 43% for13-year-olds and 33% for the 16-year-olds, were in favor of the French-speaking students.

Mathematics. Forty-eight mathematics items (38%) were identified as DIF for the 13-year-olds and 41 mathematics items (33%) for the 16-year-olds. Two of the DIF items for the 13-year-olds were at Level 3, and 3 of those for 16-year-olds were at Level 3. Twenty-seven of the DIF items were the same for the two age groups. For the 13-year-olds, the DIF items were approximately evenly distributed between the language groups, whereas more items favored the French-speaking group for the 16-year-olds. The same number of items (23) favored the French-speaking group, but this number was 49% of the DIF items for the 13-year-olds and 58% of the DIF items for the 16-year-olds.

Science. Using the L–H method, across all three science forms, 79 items (55%) were identified as DIF for the 13-year-olds and 67 items (47%) for the 16-year-olds; 10 of these DIF items were identified to be at Level 3 for each of the age groups. A smaller number of items were identified as DIF by the SIBTEST method: 54 items (38%) for the 13-year-olds and 51 items (35%) for the 16-year-olds. Yet, larger numbers of items were identified to be at Level 3. For 13-year-olds, 39 Science DIF items were identified to be at Level 3, and for 16-year-olds, this number was 32. The L–H method identified a larger number of DIF items favoring the French-speaking group than SIBTEST, (53% for 13-year-olds and 51% for 16-year-olds). Similar to the DIF patterns in reading, larger numbers of DIF items were identified by SIBTEST for the English-speaking group (54% for 13-year-olds and 55% for 16-year-olds).

TABLE 3
Number of DIF items in Reading Form A, Mathematics and Science

	13-year-olds				*16-year-olds*			
	Pro-English		*Pro-French*		*Pro-English*		*Pro-French*	
Content Area	*L–H*	*SIBTEST*	*L–H*	*SIBTEST*	*L–H*	*SIBTEST*	*L–H*	*SIBTEST*
Reading[a]	3	4	5	3	5	6	5	3
Mathematics[b]	24		23		17		23	
Science								
Form A[c]	4	1	3	3	3	2	3	3
Form B[d]	19	18	22	12	13	14	13	11
Form C[d]	14	10	17	10	17	12	18	9

Note. DIF = differential item functioning; L–H = Linn–Harnisch; SIBTEST = Simultaneous Item Bias Test.

[a]22 items. [b]125 items. [c]12 items. [d]66 items.

DIF Attributed to Adaptation Differences

To evaluate the degree to which DIF may be due to differences caused by adaptation, reviewers rated the equivalence of the two language versions of the items. The results of the DIF analyses and their corresponding judgmental review ratings are displayed in Table 4 for reading, mathematics, and science, respectively. The focus of this table is the degree to which DIF items were identified to have adaptation related differences; therefore, the judgmental review ratings of items that were not identified as DIF are not presented. All of mathematics DIF items and reading and science items that were identified as DIF by *both* DIF detection methods were included. A rating of 0 or 1 indicates no or only minor adaptation problems; a rating of 2 indicates serious adaptation problems that may lead to performance differences; a rating of 3 indicates serious adaptation problems that were expected to lead to performance differences. The information in this table is expected to be used as evidence supporting interpretations that adaptation related differences were the source of DIF. Another source of supporting evidence for this interpretation is if DIF items are replicated for both age groups. Therefore, the number of items that were identified as DIF for both age groups by both DIF detection methods is reported in the last column of the table along with their judgmental review ratings.

As summarized in Table 4, all of the reading DIF items were identified to have adaptation related differences. Among the four DIF items for the 13-year-olds, three were also identified as DIF for the 16-year-olds.

TABLE 4
Differential Item Functioning and Judgmental Review Ratings

Content Area	*Judgmental Review Rating*	*13-Year-Olds*	*16-Year-Olds*	*Common Across 13- and 16-Year-Olds*
Reading	0–1	0	0	0
	2	3	3	2
	3	1	4	1
Mathematics	0–1	30	25	17
	2	11	9	5
	3	6	6	4
Science	0–1	24	27	18
	2	17	12	10
	3	11	10	7

Note. Judgmental review rating: 0–1 - minimal or no difference in meaning between the two versions; **2** - clear differences in meaning between the two versions, but they may not necessarily lead to differences in performance between two groups; **3** - clear differences in meaning between the two versions that are expected to lead to differences in performance between two groups.

In mathematics, 17 out of 47 DIF items (36%) for 13-year-olds and 15 out of 40 DIF items (38%) for 16-year-olds were identified as having adaptation related differences. Twenty-six DIF items replicated for both age groups, 9 of which were identified to have adaptation related differences.

In the science comparisons, out of 144 items, 52 items were identified as DIF for 13-year-olds and 49 items for 16-year-olds by both methods. Thirty-five of these items replicated for both age groups. Twenty-eight of the DIF items (54%) for the 13-year-olds were interpreted to have adaptation related differences. Twenty-two of the DIF items (45%) for the 16-year-olds were interpreted to have adaptation related differences, 17 of which were the same as the items identified as DIF for the 13-year-olds.

DIF Attributed to Curricular Differences This section summarizes the degree to which DIF may be due to curricular differences by examining the relative distribution of DIF items by topic area for mathematics and science separately. Clustering of large proportions of DIF items by topic area in favor of or against a particular group was interpreted as DIF due to curricular differences. This approach for clustering of DIF items was conducted for mathematics and science for which items were categorized by topic and was not conducted for reading items which did not have such categorization. Tables 5 and 6 display the number of items identified as DIF by both L–H and SIBTEST by topic area for mathematics and science, respectively.

Mathematics DIF items by topic. There were four topic areas in mathematics: algebra and functions, measurement and geometry, numbers and operations, and data management and statistics. Each of the mathematics topic areas had a considerable number of DIF items (between 22% and 46%). While the algebra and functions and numbers and operations DIF items were evenly distributed between language groups for both 13- and 16-year-olds, for measurement and geometry and data management and statistics DIF items clustered more in favor of one

TABLE 5
The Relative Distribution of Mathematics Items Identified as Differential Item Functioning Items by Linn–Harnisch, by Curricular Topic Area

	13-Year-Olds		*16-Year-Olds*	
Curricular Topic	*Pro English*	*Pro French*	*Pro English*	*Pro French*
Algebra and functions[a]	5	7	7	8
Measurement and geometry[b]	2	8	0	8
Numbers and operations[c]	11	7	5	6
Data management and statistics[d]	6	1	5	1

[a]32 items. [b]37 items. [c]39 items. [d]17 items.

TABLE 6
The Relative Distribution of Science Identified as DIF Items
by Linn–Harnisch and Simultaneous Item Bias Test DIF Items,
by Curricular Topic Area

	13-Year-Olds		*16-Year-Olds*	
Curricular Topic	*Pro English*	*Pro French*	*Pro English*	*Pro French*
Biology[a]	4	6	3	5
Chemistry[a]	5	3	3	4
Earth[a]	2	6	2	4
Physics[a]	3	5	3	5
Nature of science[b]	5	2	6	1
Science technology in society[b]	8	3	10	3

[a]22 items. [b]28 items.

language group compared to the other. For example, 80% of the measurement and geometry DIF items for the 13-year-olds and 100% of the DIF items for the 16-year-olds in this topic area were in favor of the French-speaking examinees. Large proportions of the data management and statistics DIF items were in favor of the English-speaking-examinees. In this topic area, 86% of the DIF items for 13-year-olds and 83% of the DIF items for 16-year-olds were in favor of the English-speaking examinees. This pattern of results provided support for the interpretation that measurement and geometry DIF items in favor of the French-speaking group and data management and statistics DIF items in favor of the English-speaking group might be due to curricular differences.

Science DIF items by topic. There were six topic areas in science: biology, chemistry, earth, physics, nature of science, and science technology in society. Similar to the interpretation of mathematics DIF items, when large proportions of DIF items in a curricular topic area were in favor of one group, DIF was interpreted as due to a potential curricular difference. In science, similar to the mathematics results, all topic areas had a considerable number of DIF items (between 25% and 46%). For nature of science and science technology in society, most of the DIF items were in favor of the English-speaking group (for both 13- and 16-year-olds). Seventy-one percent and 86% (13- and 16-year-olds, respectively) of the nature of science DIF items were in favor of the English-speaking group. Seventy-three percent and 77% of the science technology in society DIF items (for 13- and 16-year-olds, respectively) also favored the English-speaking group. DIF items were more evenly distributed on the other four topic areas. This pattern of results suggests that DIF items favoring the English-speaking group in nature of science and science technology in society may be due to curricular differences.

SUMMARY AND DISCUSSION

This article examined the degree of comparability and sources of incomparability of the English and French versions of SAIP tests across three content areas. The results point to substantial psychometric differences between the two language versions of tests at the item level. In reading comparisons, approximately 18% to 31% of the items were identified as DIF by both detection methods across the two age groups. In mathematics, only the L–H DIF detection method was used. Based on this method, 32% to 37% of the items displayed DIF for the two language groups across the two age groups. In the science comparisons, 32% to 36% of the items were identified as DIF by both methods for 13- and 16-year-olds. These results reveal that a relatively large number of DIF items were identified, a finding reported by other researchers in the area of test translation and adaptation (e.g., Allalouf, Hambleton, & Sireci, 1999; Ercikan, 1999; Ercikan & McCreith, 2002; Gierl, Rogers, & Klinger, 1999; Gierl & Khaliq, 2001).

Some differences were observed between the DIF detection patterns of L–H and SIBTEST that are worth noting and will require further investigation. First, the L–H DIF detection method identified larger numbers of DIF items. Second, SIBTEST identified more DIF items for English-speaking examinees in both reading and science. Third, SIBTEST identified much greater numbers of Level 3 DIF. In fact, in three out of four tests that it was used to identify DIF, it identified more Level 3 DIF items than Level 2 DIF items. Further analyses need to be conducted to examine the relative sensitivity of the two methods to DIF, whether either of the methods are biased in detecting DIF in one direction or the other and classification of different levels of DIF.

Reviews of adaptations identified 36% to 100% of DIF items to have differences due to adaptations in the three content areas and across two age groups. In addition to reviews by bilingual translators, two other approaches were taken to help interpret sources of DIF. First, DIF analyses for two age groups were replicated. Second, curricular differences as a potential source of DIF were investigated by examining grouping of DIF by curricular area. Even though DIF analyses in the two age group comparisons is not expected to be identical, if DIF is indeed due to adaptation effects, it is reasonable to expect that adaptation differences would lead to differential functioning in both of the comparisons. As can be seen from the results, DIF items associated with adaptation differences were, in fact, replicated in both comparisons in 75% of the cases in reading, 71% of the cases in mathematics, and 61% of the cases in science.

Clustering of DIF items by topic provided curricular differences as source of DIF only for small portions of the DIF items, approximately 25% to 17% of the mathematics DIF items and 27% to 33% of the science DIF items. Some of the DIF items that were attributed to curricular differences were also identified to have adaptation related differences. The number of items that were attributed to adapta-

tion, curricular, both of these sources, and those DIF items sources of which could not be identified are presented in Table 7. As can be seen in this table, 7 out of 17 mathematics DIF items for 13-year-olds and 4 out of 22 mathematics DIF items for 16-year-olds that were interpreted to have adaptation related differences were also attributed to curricular differences. These DIF items may indeed be due to both adaptation and curricular differences. It is also possible that the degree of effect of these differences on DIF and whether these are in fact the sources of DIF cannot be determined without further investigation using experimental designs and examination of student cognitive processes.

There were large portions of DIF items that could not be attributed to adaptation-related or curricular differences, which is another finding reported by some researchers in this area (e.g., Ercikan, 2002; Gierl et al., 1999; Gierl & Khaliq, 2001). For example, for the 13-year-olds, 42% of the mathematics DIF items and 41% of the science DIF items were not linked to the two sources of DIF examined in this study. In addition to the possibility that the procedures used in this study did not identify all DIF items that were associated with adaptation or curricular differences, several factors that were not considered in this article could be potential explanations for DIF. Among these include differences in instruction methods, cultural differences, and limitations in definitions of topics.

The findings of this study further highlight that comparability of language versions of assessments cannot be assumed, and empirical examinations of comparability is essential to validity of interpretations. In addition, multiple sources of incomparability need to be considered. Further research needs to investigate the range of differences in examinees and tests that may contribute to differences in constructs assessed by items and tests and ways of disentangling these sources of differences.

TABLE 7
Numbers and Percentages of Differential Item Functioning Items with Adaptation and Curricular Interpretations

		Adaptation[a]		*Curriculum*		*Adaptation and Curriculum*		*Uninterpreted*		
Age	*Content Area*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*Total Number*
13-year-olds	Reading	4 (3)	100	—		—		0		4
	Mathematics	17 (9)	36	14	30	7	15	23	49	47
	Science	28 (17)	57	13	27	4	8	12	24	49
16-year-olds	Reading	3 (3)	43	—		—		4	57	7
	Mathematics	15 (9)	38	13	33	5	13	17	43	40
	Science	22 (17)	45	16	33	4	8	15	31	49

[a]Numbers in parentheses indicate number of items that were identified as differential item functioning for both age groups.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement, 36,* 185–198.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Rep. No. 88–2). New York: College Entrance Examination Board.

Bock, R. D. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Council of Ministers of Education, Canada. (1997). *SAIP mathematics 1997—The public report.* Retrieved July 13, 2004, from http://www.cmec.ca/saip/math97/index.stm

Council of Ministers of Education, Canada. (1998). *SAIP reading and writing 1998—The public report.* Retrieved July 13, 2004, from http://www.cmec.ca/saip/rw98le/pages/tablee.stm

Council of Ministers of Education, Canada. (1999). *SAIP science 1999—The public report.* Retrieved July 13, 2004, from http://www.cmec.ca/saip/sci96/index.stm

Council of Ministers of Education, Canada (2000). *Report on science assessment*, *School Achievement Indicators Program, 1999*. Toronto, Ontario, Canada: Author.

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

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

Ercikan, K. (1999, April). *Translation DIF on TIMSS*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Ercikan, K. & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In D. Robitaille & A. Beaton (Eds.), *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391–407). Dordrecht, the Netherlands: Kluwer.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 4*, 199–215.

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

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment, 9*, 57–68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–224.

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

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

Lord, F. M. (1952). *A theory of test scores*. (Psychometrika Monograph No. 7). Richmond, VA: William Byrd Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Muraki, E. (1992). A generalized partial credit model: Application of EMalgorithm. Applied Psychological Measurement, *16,* 159–176.

O'Neil, K. A., & McPeek, W. M., (1993). In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel–Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215–230.

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 35*, 229–259.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). *Adapting credentialing examinations for international uses*. (Laboratory of Psychometric and Evaluative Research Report No. 329). Amherst: University of Massachusetts, School of Education.

Shealy, R., & Stout, W.F. (1993). An item response theory model for test bias. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ:Lawrence Erlbaum Associates, Inc.

van der Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology, 47*, 263–279.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–214.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55–66.