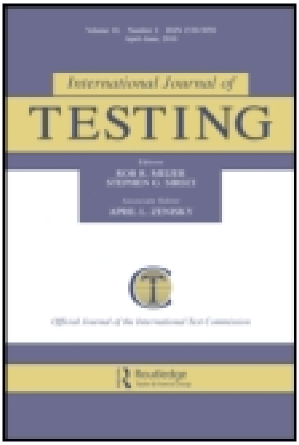

International Journal of Testing

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hijt20

Disentangling Sources of Differential Item Functioning in Multilanguage Assessments

Kadriye Ercikan

Available online: 22 Jun 2011

To cite this article: Kadriye Ercikan (2002): Disentangling Sources of Differential Item Functioning in Multilanguage Assessments, International Journal of Testing, 2:3-4, 199-215

To link to this article: http://dx.doi.org/10.1080/15305058.2002.9669493

INTERNATIONAL JOURNAL OF TESTING, *2*(3&4), 199–215
Copyright © 2002, Lawrence Erlbaum Associates, Inc.

Disentangling Sources of Differential Item Functioning in Multilanguage Assessments

Kadriye Ercikan
Faculty of Education
University of British Columbia

This article describes and discusses strategies used in disentangling sources of differential item functioning (DIF) in multilanguage assessments where 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 multiple groups, and (c) examination of the distribution of DIF items by topic. Twenty-seven percent of the mathematics DIF items and 37% of the science DIF items were interpreted to be due to adaptation-related differences based on judgmental reviews. Most of these interpretations were also supported by the cross-validation analyses. Clustering of DIF items by topic provided curricular differences as interpretation for DIF only for small portions of the DIF items, approximately 23% of the mathematics DIF items and 13% of the science DIF items.

Multilanguage assessments are defined as assessments that are administered in more than one language. A basic concern of multilanguage assessments is whether performance on test items is comparable when items are adapted to different languages. Previous research has identified that differences due to adaptations of tests in multiple languages can cause problems in comparability and equivalence of items in different languages (Allalouf, Hambleton, & Sireci, 1999; Ercikan, 1998; Ercikan & McCreith, 2002; Gierl & Khaliq, 2001; Hambleton, 1993). In international assessments, where multilanguage versions of assessments are administered in different countries, there are several other factors that might affect item equiva-

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

lence. These, for example, include cultural and curriculum differences in these countries. 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). For example, inclusion of a map in an item that might not require previous knowledge of content of the map might make the item easier or more difficult for some examinees, depending on their levels of familiarity with locations on the map. 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 had 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 undoubtedly have differential coverage in different countries, and this differential coverage can lead to differential response patterns and difficulty levels, independent of any problems due to adaptation.

One of the most commonly used methods for examining equivalence or comparability of test items for different groups is statistical techniques used for identifying differential item functioning (DIF). The statistical methods used for identifying DIF are designed to identify items that result in differing psychometric properties for the comparison groups, such as males and females. Van der Linden (1998) discussed use of DIF identification methods for identifying items that are functioning differentially for different language groups due to adaptations. He stated that:

> [T]his might even be one of the very few cases where this technique can be applied meaningfully …. It does make much sense to test the statistical null hypothesis of no difference between these versions against the alternative hypothesis of some of the versions functioning differentially. (p. 573)

Several researchers have applied DIF detection procedures for identifying test items that are functioning differentially in different languages (Allalouf et al., 1999; Ercikan, 1998; Gierl & Khaliq, 2001; Sireci, Fitzgerald, & Xing, 1998). One of the challenges in using DIF for examining comparability of test items is the identification of potential causes of DIF. Several researchers have investigated potential causes of DIF for gender and ethnic groups (O'Neil & McPeek, 1993; Scheunemann & Geritz, 1990; Schmitt & Dorans, 1990; Zwick & Ercikan, 1989). Hambleton and Jones (1994) discussed the complexity of factors that might affect DIF and stated:

> An item may be functioning differentially if it contains content or language that is differentially familiar to subgroups of examinees, or if the item structure or format is

> differentially familiar to subgroups of examinees, or if the item structure or format is differentially difficult for subgroups of examinees. (p. 24)

The existence of multiple factors affecting equivalence of test items in these comparisons makes it difficult to identify sources of DIF. There have been mixed levels of success in identifying sources of DIF. Engelhard, Hansche, and Rutledge (1990) found that the agreement between judgmental reviews and the empirical DIF levels was generally not beyond what would be expected by chance. Allalouf et al. (1999), on the other hand, reported that reviewers identified sources of DIF in more than 80% of the items that functioned differently across Hebrew and Russian versions of the test items. Gierl and Khaliq (2001) reported 62% to 88% success levels in identifying sources of mathematics and social studies DIF items using judgmental reviews. Much poorer results were reported by Sireci et al. (1998) in identifying sources of DIF in English–German versions of a test using an electronic reviewing process. They found that translators' impressions were generally inconsistent with the DIF analyses. Ercikan (1998) reported that 44% of DIF items in an international assessment were linked to adaptation-related differences and the sources of DIF for the rest of the items were not identified.

Certain methodological aspects of judgmental reviews, such as the proportion of DIF items in the pool of reviewed items and the standardization of the training of reviewers and the reviewing criteria, are expected to affect the success rates for identifying sources of DIF. In the studies reviewed earlier, Sireci et al. (1998) identified lack of possibility of standardization of the rating process as an important reason for inconsistency between the judgmental reviews and DIF identification. Gierl and Khaliq (2001) used a consensus-building model wherein the reviewers worked as a group and focused on standardizing interpretations and ratings across reviewers, which may have contributed to high success rates of explaining DIF. On the other hand, a large proportion (70%) of the items in the Allalouf et al. (1999) study were DIF, which might have contributed to high success rates for reviewers in identifying DIF items. In contrast, only 26% of the items were identified as DIF in the Ercikan (1998) study. In addition to differences in the degree of standardization and varying proportions of DIF items in the item pools, there are other fundamental differences among the review processes that affect the degree to which judgmental reviews can explain or predict DIF. These include (a) whether the reviewers have knowledge about the DIF status of items, (b) whether single or multiple versions of tests are reviewed, and (c) number of factors as sources of DIF. In some of the studies discussed earlier, reviewers did not have knowledge of the DIF status of items, and they were trying to predict DIF based on comparability of test items in different language versions. This was true for all the studies already discussed except for the Gierl and Khaliq (2001) study. In the Gierl and Khaliq study, the reviewers knew which items were differentially functioning and were trying to identify and classify sources of DIF. In the judgmental review process, when the review of items is limited to DIF items, the reviewing process is reduced to identifying direction of DIF and identification of differences between language versions that might cause DIF. The success rate for identifying direction of DIF in such a review is at least 50%, which is the success rate due to chance.

The second difference among the studies already discussed is whether a single or multiple versions of tests are being reviewed. Engelhard et al. (1990) examined the same version of the test items for the comparison groups, whereas the other studies were examining different language versions of test items. In multilanguage versions of tests, reviews of comparability of tests focus on comparability of format, content, and language, and target 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. Previous research indicated that low success rates in identifying sources of DIF for gender and ethnic groups were common (Scheuneman, 1987).

The third difference among the studies discussed is that different numbers of factors might be affecting DIF. Different factors might be at play even in multilanguage assessments depending on whether a test is an achievement test or a licensure, test and whether the test is administered in one country or multiple countries. For example, for achievement tests, performance 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 might not be as important sources of DIF. When tests are administered in multiple countries, in addition to curricular differences, cultural differences are expected to be larger between countries. Therefore, cultural differences are expected to be more important sources of DIF.

The purpose of this article is to describe and discuss strategies used in disentangling sources of DIF in multilanguage assessments where 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 multiple groups, and (c) examination of the distribution of DIF items by topic. Other major factors that might affect DIF, such as cultural and instructional differences, are not examined, so interpretations of DIF are not exhaustive. Rather, an evidence-gathering approach is taken, whereby analyses are conducted to gather evidence that might provide support for or against the following hypotheses.

HYPOTHESIS 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 two additional comparisons.

HYPOTHESIS 2: DIF IS DUE TO CURRICULAR DIFFERENCES

Clustering of DIF items that are all in favor of or against the same group in a 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 might be due to curricular differences in relation to this topic.

METHOD

Instrument

This study examined adaptation effects in the 1995 Third International Mathematics and Science Study (TIMSS) assessment of 13-year-old students' (Population 2) mathematics and science knowledge in 45 countries. The tests consisted of a pool of mathematics and science items that were matrix-sampled across eight booklets. The pool of items was divided into 26 sets of items that were then arranged in various ways to make up eight test booklets, each containing seven item clusters. One cluster, the core cluster, appears in each booklet. Seven focus clusters appear in three of the eight booklets. There are also 12 breadth clusters, each of which appears in just one test booklet. Finally, there are eight free-response clusters, each of which appears in two booklets (Martin, 1996).

Data

The study focused on comparability of English and French versions of items in the TIMSS and used data from test administrations in Canada (in English and French), England, France, and the United States. The selection of these countries allowed for examination of comparability of English and French versions of the test items in the Canadian administration and cross-validating the findings in two other comparisons (i.e., England–France and United States–France) where the same versions of items were administered. The sample sizes and number of items for these assessments are presented in Table 1.

TABLE 1
Sample Sizes and Number of Items Used in Comparisons

		Number of Items	
Country	*N*	*Mathematics*	*Science*
Canada		156	140
English speaking	5,445		
French speaking	2,925		
England–France		156	139
England	3,572		
France	6,002		
United States–France		154	139
United States	10,945		
France	6,002		

Identification of DIF

DIF was identified using a DIF detection procedure described by Linn and Harnisch (1981) and developed using an item response theory (IRT) based approach (CTB/McGraw-Hill, 1991). This procedure computes the observed and expected mean response (expected and observed *p* values) and the difference between them (observed minus predicted, p_{diff}) for each item by deciles of the specified group. The expected values are computed using the parameter estimates obtained from the entire sample and the theta estimates (ability estimates) for the members of the specified subgroup. Based on the difference between expected and observed *p* values, a *Z* statistic is calculated for each decile and an average *Z* statistic for the item is computed for identifying the degree of DIF. The DIF status of an item is determined by the statistical significance of the *Z* statistic. To distinguish between different levels of DIF, CTB/McGraw-Hill developed a set of rules using a systematic analysis of different levels of *Z* statistic and the degree of difference between the item characteristic curves for each of the comparison groups. These rules are summarized in Table 2. A negative difference implies bias against the subgroup. Items are flagged as biased for or against the specified subgroup according to the following rule: An item is classified as Level 3 DIF if the absolute value of the obtained minus the expected *p* value is greater than or equal to 0.10, and also the corresponding absolute *Z* value is $|Z| > 2.58$. If $|Z| \geq 2.58$ but the absolute value of the *p* value difference is less than 0.10, 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. There are some advantages to identifying DIF using the IRT-based Linn–Harnisch (LH) method with the TIMSS test design that involves eight overlapping test booklets. Popularly used DIF detection procedures such as logistic regression (Swaminathan & Rogers, 1990) or Mantel–Haenzsel (Holland & Thayer, 1986) would require identification of DIF separately by booklet, which means that multiple analyses are conducted for items that are in more than one test booklet. IRT allows us to combine items across all overlapping test booklets and obtain comparable results for items across test booklets. In this case, the IRT-based LH procedure allows us to combine items across all test booklets. This results in a single DIF analysis for each item. In the LH procedure, multiple-choice items were calibrated using the three-parameter logistic model (Lord, 1980) and the open-ended items were calibrated using the two-parameter partial credit model (Yen, 1993). A simultaneous calibration of multiple-choice and open-ended items and identification of DIF were conducted using PARDUX (CTB/McGraw-Hill, 1991).

In previous research, Ercikan (1999) compared the identification of DIF by LH to that of the logistic regression method (Swaminathan & Rogers, 1990) using the same TIMSS data used in this study and conducting analyses separately for each booklet. The DIF identification was consistent for 91% of the mathematics items and 89% of the science items. The limitation of use of an IRT-based method for DIF detection in TIMSS design is that the sample sizes can be too small for obtaining stable results. Combining items across booklets results in larger sample sizes for each test item and such an analysis design was preferred in this study.

Judgmental Reviews

Four translators bilingual in French and English, for potential adaptation-related differences, examined the French and English versions of all items administered in Canada. The translators did not have any knowledge of which items displayed DIF. If any differences were identified in the review process, each translator who would rate the degree to which there were differences in meaning due to adaptations evaluated the two versions of items. Items were assigned ratings between 0 and 3. A

TABLE 2
Statistical Rules For Identifying Three Levels of DIF

DIF Level	*Rule*	*Implication*
Level 1	$\lvert Z \rvert < 2.58$	No DIF
Level 2	$\lvert Z \rvert \geq 2.58$ and $\lvert p_{diff} \rvert < 0.10$	Moderate DIF
Level 3	$\lvert Z \rvert > 2.58$ and $\lvert p_{diff} \rvert \geq 0.10$	High DIF

Note. DIF = differential item functioning.

rating of 0 indicated there were no differences in meaning in the adaptation of the item. Ratings between 1 and 3 were assigned to items in which there were adaptation problems identified between the two versions, indicating the degree of impact on the meaning of the item. For example, a rating of 1 was assigned when there were minimal differences in meaning due to adaptations. Items were assigned a rating of 2 if there were clear differences between the two versions, but they were not expected to lead to differences in examinee performance for the two groups. A rating of 3 was assigned if the adaptation problems were expected to lead to differences in comparability of performances for the two groups. An average of ratings from the four reviewers was used as the judgmental review rating for that item.

Analyses

Analyses were intended to determine (a) whether there was differential functioning between English and French versions of items, and (b) whether differential functioning could be attributed to curricular or adaptation differences. Pairwise DIF analyses were conducted separately for the Canadian English and French item versions, for England and France, and for the United States and France. These analyses provided information about the degree of DIF in the English and French versions of items. If DIF in the Canadian analysis was replicated in the other two comparisons in the same direction, it provided support for the interpretation that the differences might be due to adaptations rather than cultural or curricular differences. If, on the other hand, DIF was in the same direction (e.g., all in favor of French-speaking students) and clustered by topic, then there was support for the interpretation that DIF might be due to curricular differences.

RESULTS

There were approximately 3,000 to 11,000 cases in each comparison group. Items were combined across test booklets and analyses were conducted separately for mathematics and science. This resulted in 139 to 156 items for each comparison (see Table 1). The results of the analyses examining degree of DIF in English and French versions of items that provided evidence of whether DIF was due to adaptation effects or curricular effects are presented next.

DIF in English and French Versions of Items

The results of the DIF detection procedure are summarized in Table 3 for mathematics and science items. Out of 156 items, 22 items were identified as DIF in the English and French versions of the mathematics portion of the TIMSS that were administered in Canada. Forty-one percent of mathematics DIF items functioned

TABLE 3
Numbers DIF Items in the Canadian Analyses

	Pro-English		*Pro-French*	
Subject	*Level 2 DIF*	*Level 3 DIF*	*Level 2 DIF*	*Level 3 DIF*
Mathematics	11	2	8	1
Science	16	2	30	4

Note. DIF = differential item functioning.

in favor of the French-speaking group. More items were identified as DIF in science than in mathematics. There were 52 DIF science items out of 140 items, 65% of which functioned in favor of the French-speaking group.

Replication of DIF Items in the Three Comparisons

The focus of these analyses was the cross-validation of DIF findings in the Canadian comparisons in the other two comparisons, namely, the England–France and United States–France comparisons.

Mathematics DIF items. The numbers of mathematics items identified as DIF in the Canadian analyses and those that were replicated in the England–France and United States–France comparisons are presented in Table 4. In mathematics, out of 13 DIF items in favor of the English-speaking group in the Canadian comparison, 6 were replicated in the England–France comparison and 6 were replicated in the U.S.–France comparison. Two of the DIF items were replicated in all three comparisons. Out of 9 mathematics items in favor of the French-speaking group in the Canadian comparison, 3 items were identified to be in favor of the French speakers in the England–France comparison and 6 were in favor of the French speakers in the U.S.–France comparison. Two of the items were identified as DIF in all three comparisons.

Science DIF items. The numbers of science items identified as DIF in the Canadian analyses and those that were replicated in the England–France and U.S.–France comparisons are presented in Table 5. In science, out of 18 DIF items in favor of the English-speaking group in the Canadian comparison 7 were also identified as DIF favoring English speakers in the England–France comparison, and 11 were identified as DIF in the U.S.–France comparison. There were six items identified as favoring the English-speaking groups in all three comparisons. Out of 34 science items in favor of the French-speaking group in the Canadian comparison, 13 were in favor of French speakers in the England–France compari-

TABLE 4
Replication of Mathematics DIF Items in the Canadian Comparison in the England–France and United States–France Comparisons

		Canada				*England–France*				*United States–France*			
		Pro-English		*Pro-French*		*Pro-English*		*Pro-French*		*Pro-English*		*Pro-French*	
		L2	*L3*	*L2*	*L3*	*L2*	*L3*	*L2*	*L3*	*L2*	*L3*	*L2*	*L3*
Pro-English	L2	11	—	—	—	4	—	1	—	4	1	1	2
	L3	—	2	—	—	2	—	—	—	—	1	—	—
Pro-French	L2	—	—	8	—	1	—	2	1	—	—	2	3
	L3	—	—	—	1	—	—	—	—	—	—	—	1

Note. L = level of DIF.

TABLE 5
Replication of Science DIF Items in the Canadian Comparison In the England–France and United States–France Comparisons

		Canada				*England–France*				*United States–France*			
		Pro-English		*Pro-French*		*Pro-English*		*Pro-French*		*Pro-English*		*Pro-French*	
		L2	*L3*	*L2*	*L3*	*L2*	*L3*	*L2*	*L3*	*L2*	*L3*	*L2*	*L3*
Pro-English	L2	16	—	—	—	7	—	1	1	7	3	2	1
	L3	—	2	—	—	—	—	—	—	1	—	—	—
Pro-French	L2	—	—	30	—	3	—	10	2	4	2	11	5
	L3	—	—	—	4	—	—	1	—	—	—	—	1

Note. L = level of DIF.

son and 17 were in favor of the French speakers in the U.S.–France comparison. Ten of these items were common across the three comparisons.

Judgmental Reviews

Reviews of all the items were conducted prior to examining the degree to which adaptation-related problems were associated with differential functioning. The translators identified adaptation-related problems in 64 out of 296 (156 mathematics, 140 science) items. Twenty-five (approximately 39%) of these items were identified as differentially functioning in the Canadian English and French comparisons. The results of the DIF analyses and their corresponding judgmental review ratings are displayed in Tables 6 and 7 for mathematics and science, respectively. Fifteen out of 22 of the mathematics DIF items had judgmental review ratings of 1 or lower, indicating minor to no differences between French and English versions of the Canadian assessments. In mathematics, 6 DIF items were identified to have adaptation differences. All of these items had judgmental review ratings of 2. Two of these items had high DIF levels and 4 had moderate DIF levels.

Twenty-nine out of 52 science DIF items had judgmental review ratings of 1 or lower, indicating little to no difference between the adapted versions. Nineteen science DIF items were identified to have adaptation-related differences. Three of these items had high DIF levels; 1 of these items had a judgmental review rating of 2 and the other 2 items had review ratings of 3. There were 17 moderate-level DIF items, 10 of which had judgmental review ratings of 2 and 7 with review ratings of 3.

There were several observations made during the judgmental reviews of this study that linked different aspects of adaptations of tests to DIF. First was that the association between the judgmental review ratings and DIF levels were moderate. Even though the differences may seem small in adapted versions of a test, these differences were associated with high levels of DIF. On the other hand, large differences in meaning were not necessarily associated with high levels of DIF. Approximately half of the items that were identified by the bilingual translators as having serious adaptation problems displayed DIF as identified by the statistical

TABLE 6
Mathematics DIF Items and Judgmental Review Ratings

	Differential Item Functioning Level	
Judgmental Review Rating	*2*	*3*
0–1	15	1
2	4	2
3	0	0

TABLE 7
Science DIF Items and Judgmental Review Ratings

	Differential Item Functioning Level	
Judgmental Review Rating	*2*	*3*
0–1	29	4
2	10	0
3	7	2

procedure. Approximately 26% of the mathematics and 38% of the science DIF items had adaptation-related explanations.

There were three problems in adaptations that were interpreted as leading to DIF by the judgmental reviews. First is the inadequate adaptation of key words. Key words could be important information that was expected to guide examinee thinking processes. Inappropriate adaptations of words did not necessarily lead to differential functioning of items for the two groups. However, it was observed that exclusion or inappropriate adaptation of key words led to DIF. Second, adaptation problems were related to the commonness of vocabulary in a given context. Third, problems were related to the look and formatting of the item.

Evidence Supporting Adaptation Effects as Explanations for DIF

Examination of replication of DIF in three comparisons resulted in supporting adaptation-related interpretation of DIF in the Canadian comparisons (see Table 8). In mathematics, two of the six DIF items that were interpreted as adaptation related were replicated in all three comparisons. The other four items were replicated in two of the three comparisons. In science, 9 of the 19 DIF items that were interpreted to be related to adaptation differences were replicated in all three comparisons; another 5 were replicated in two of the three comparisons. In summary, all of the six mathematics DIF items that were associated with adaptation-related differences were replicated in at least two of the comparisons. In science, 14 of the 19 DIF items that were associated with adaptation-related differences were replicated in at least two of the comparisons.

Evidence Supporting Curricular Differences as Explanations for DIF

This section summarizes the degree to which DIF in the Canadian comparisons may be due to curricular differences by examining the relative distribution of DIF items by topic area for mathematics and science separately (see Tables 9 and 10).

TABLE 8
Number of Adaptation Related DIF Items Replicated in Three or Two Comparisons

	Adaptation DIF	*Replicated in Three Comparisons*	*Replicated in Two Comparisons*
Mathematics	6	2	4
Science	9	9	5

Note. DIF = differential item functioning.

TABLE 9
The Relative Distribution Of Mathematics DIF Items By Topic Area

	Level 2		*Level 3*	
Topic	*Pro-English*	*Pro-French*	*Pro-English*	*Pro-French*
Algebra (29 items)	2	1	1	—
Data representation, analysis, and probability (20 items)	1	—	—	—
Fractions and number sense (52 items)	4	2	1	—
Geometry (23 items)	1	5	—	—
Measurement (21 items)	3	—	—	1
Proportionality (12 items)	—	—	—	—

TABLE 10
The Relative Distribution of Science DIF Items by Topic Area

	Level 2		*Level 3*	
Topic	*Pro-English*	*Pro-French*	*Pro-English*	*Pro-French*
Chemistry (21 items)	1	3	—	—
Earth science (23 items)	1	8	—	2
Environmental issues and nature of science (16 items)	3	5	1	—
Life science (41 items)	5	5	1	1
Physics (40 items)	6	9	—	1

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.

Mathematics DIF items by topic. There were six area topics in mathematics, namely algebra, data representation and analysis and probability, fractions and number sense, geometry, measurement, and proportionality. Five of the mathematics topics had less than 15% of the items on that topic identified as DIF. However 26% (6 of 23) of the Geometry items were identified as DIF. Five of these items, 83% of DIF items on this topic, were in favor of the French-speaking group, which provides support for the interpretation that DIF in these five items may be due to curricular differences.

Science DIF items by topic. There were five topic areas in science: chemistry, earth science, environmental issues and nature of science, life science, and physics. Similar to the interpretation of mathematics DIF items, when large proportions of DIF items were in favor of one group, DIF was interpreted as due to curricular differences. Seventy-five percent of the chemistry DIF items and 91% of the earth science DIF items were in favor of the French-speaking group. DIF items were more evenly distributed between the two groups in the other three topic areas. Given the small number of DIF items, the strongest evidence for curricular differences between English- and French-speaking groups as a source of DIF was on earth science items. Ten out of 11 earth science items were in favor of the French-speaking group.

SUMMARY AND DISCUSSION

This study reported analyses to gather evidence for or against two hypotheses: (a) DIF is due to adaptation effects, and (b) DIF is due to curricular differences. The numbers of DIF items identified with adaptation or curriculum-related interpretation in the Canadian comparison are presented in Table 11. In mathematics, 27% of the DIF items were interpreted to be associated with adaptation-related differences. This interpretation was supported by both the adaptation reviews and by the cross-validation analyses. In science, 37% of the DIF items were interpreted to be due to adaptation-related differences based on the judgmental review. Most of these interpretations, 14 out of 19, were also supported by the cross-validation analyses.

Even though DIF analyses in the three comparisons are 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 all three comparisons. As can be seen from the results, DIF items that were interpreted to be associated with adaptation differences were in fact replicated in at least two of the comparisons in 74% of the cases. The four items that were not replicated in the two cross-validations were all very low DIF, and barely made the Level 2 DIF criteria.

TABLE 11
Number and Percentage of DIF Items in the Canadian Comparison With Adaptation and Curricular Interpretations

Mathematics DIF Items				*Science DIF Items*			
Adaptation	*Curriculum*	*Uninterpreted*	*Total*	*Adaptation*	*Curriculum*	*Uninterpreted*	*Total*
6 (2, 4)	5	11	22	19 (9, 5)	7	26	52
27%	23%	50%		37%	13%	50%	

Note. The first number in parentheses refers to the number of items that were replicated in all three comparisons, and the second number refers to the number of DIF items replicated in two comparisons. DIF = differential item functioning.

Therefore, even if adaptation-related differences were the source of DIF in these items, their effects were low in the Canadian comparison and were not replicated in the other two comparisons.

There might be other reasons why some of the adaptation-related DIF items were not replicated in all three comparisons. DIF captures similarity of constructs being assessed in the two comparison groups. This similarity is expected to be somewhat different in each of the comparisons (Canadian, England–France, and United States–France) due to differing degrees of similarities in curriculum, instruction, and culture. In addition, small differences in adapted versions of tests might lead to DIF in some populations and not others due to differences in ability distributions as well as differential effects of item language, content, or format on examinee performance in that population. For example, even though the same versions of tests, English and French, are being compared, the effects of item language on the French-speaking groups are potentially different in France and in Canada.

Clustering of DIF items by topic provided curricular differences as interpretation for DIF only for small portions of the DIF items—approximately 23% of the mathematics DIF items and 13% of the science DIF items. There were large portions of DIF items that could not be attributed to adaptation-related or curricular differences. These were 50% of the mathematics DIF items and 50% of the science DIF items. 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.

As a final note, I would like to highlight that (a) there are multiple sources of DIF, and (b) depending on the type and purpose of the assessment and comparison groups, the meaning and sources of DIF may vary. In multilanguage assessments, equivalence of items for the comparison groups is further challenged by the multiple language versions of tests. Even though identification of sources of DIF due to adaptation-related differences is relatively easier, other factors that might affect comparability such as differential familiarity with context of items, curricular differences, and cultural differences need to be considered as potential sources of DIF. It is also important to keep in mind that the effects of these factors on DIF will be different depending on whether the test is a psychological test, achievement test, or licensure test. In identifying sources of DIF using judgmental reviews, the interpretations and efforts to identify sources of DIF can be speculative. This speculative nature is further magnified if the reviewers have knowledge about which items are differentially functioning. Multiple sources need to be considered in examining sources of DIF and judgmental review processes that focus on one source should not be expected to provide explanations for all items that are differentially functioning.

REFERENCES

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

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

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347–360.

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

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

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

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

Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly, 1,* 21–36.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel–Haenzsel procedure* (Tech. Rep. No. 86–69). Princeton, NJ: Educational Testing Service.

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

Martin, M. O. (1996). Third international mathematics and science study: An overview. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study: Technical report, Vol. I: Design and development*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement, Boston College.

O'Neil, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 3,* 97–118.

Scheuneman, J. D., & Geritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109–131.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27,* 67–81.

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

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

Van der Linden, W. J. (1998). A discussion of some methodological issues in international assessments. *International Journal of Educational Research, 29,* 569–577.

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, 2,* 55–66.