Curriculum, Translation, and Differential Functioning of Measurement and Geometry Items

Barnabas C. Emenogu

Ruth A. Childs

A test item exhibits differential item functioning (DIF) if students with the same ability find it differentially difficult. When the item is administered in French and English, differences in language difficulty and meaning are the most likely explanations. However, curriculum differences may also contribute to DIF. The responses of Ontario students to Measurement and Geometry items from the content subtest of the 2001 School Achievement Indicators Program Mathematics Assessment were analyzed using item response theory-based procedures. DIF between the French and English versions was investigated. Attempts to interpret the DIF found in terms of translation and curriculum influences were partially successful. Alternative explanations and suggestions for additional research are provided.

Keywords: differential item functioning, curriculum, translation, mathematics assessment, large-scale assessment

Un item de test donne lieu à un fonctionnement différencié d'item (FDI) si des élèves ayant les mêmes aptitudes le trouvent difficile de façon différente. Lorsqu'un item est administré en français et en anglais, les différences quant aux difficultés de la langue et au sens sont les explications les plus vraisemblables. Or, les différences d'ordre curriculaire peuvent aussi contribuer au fonctionnement différencié des items. Les réponses des élèves ontariens à des items de mesure et de géométrie tirés du sous-test de contenu du Programme d'indicateurs du programme scolaire 2001 pour l'évaluation en mathématiques ont été analysées à l'aide de procédures fondées sur la théorie de la réponse aux items. Le FDI entre les versions française et anglaise a été étudié. Les tentatives d'interpréter le FDI en termes de traduction et de programmes ont été partiellement couronnées de succès. D'autres explications et des suggestions de recherches complémentaires sont données.

Mots clés: fonctionnement différencié des items, curriculum, traduction d'instruments, évaluation en mathématiques, évaluation à grande échelle.

Two students with the same level of mathematics understanding should have equal probabilities of answering a mathematics test item correctly. If their probabilities are different, the item is said to exhibit differential item functioning (DIF). Understanding the extent to which items function

differently across groups of students and explaining these differences has motivated much recent psychometric literature (e.g., Allalouf, Hambleton, & Sireci, 1999; Gierl & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999; Swanson, Clauser, Case, Nungester, & Featherman, 2002; Zwick, Thayer, & Lewis 2000). Studies have compared the functioning of items for females and males, for students of different ethnicities or cultural backgrounds, and for students taking tests in different languages.

In studies of items administered in two languages, explanations for DIF are typically sought in terms of differences in language difficulty and meaning. Guidelines for translating tests have been suggested (e.g., van de Vijver & Hambleton, 1996) to minimize translation DIF. Although language differences are the most obvious explanation for DIF between translated test forms, other explanations are also possible. For example, school curricula may differ across languages, students may be taught to solve mathematics problems using different methods, or schools may differ in the availability of textbooks or other resources.

DIF is a particular concern for tests such as those in the School Achievement Indicators Program (SAIP) because of the diversity of educational jurisdictions in which they are administered. In the SAIP, French and English versions of each test are administered to samples of 13- and 16-year-olds across Canada. Recent analyses by Boiteau, Bertrand, and St-Onge (2002); Ercikan, Gierl, McCreith, Puhan, and Koh (2002); and Koh and Ercikan (2002) found evidence of DIF between French and English versions of SAIP mathematics, science, and language assessments. Some of these differences may be due to differences in precise meaning or in vocabulary difficulty between the two language versions (translation DIF). It is also possible that some of the differences are due to differences in curricula across populations taking the tests in French and in English. Other possible factors include curriculum differences between provinces, which are complicated by the presence of both French- and English-language schools in some provinces.

To control for interprovincial differences in curricula, this study focused on the responses of Ontario students. Ontario presents a unique opportunity for exploring the impact of language and curriculum differences, both because the mathematics curriculum in Ontario has recently been replaced, so that the 13- and 16-year-old students participating in the assessment were studying under different curricula, and because differences of curricula and resources exist between Ontario's French- and English-language schools. This study explored the possible sources of DIF for the Measurement and Geometry subset of items from the 2001 SAIP Mathematics Assessment for students in Ontario schools who responded to those items.

Differential Item Functioning

Dorans and Holland (1993) defined DIF as a psychometric difference between groups that are matched on the ability or the achievement measured by an item. That is, an item exhibits DIF if it provides a consistent advantage or disadvantage to members of a group, not because of differences in the trait of interest, but because of differences in other traits or because different versions (e.g., translations) of an item measure different traits. More simply, when examinees in different groups have different probabilities of answering an item correctly after controlling for overall ability, the item is said to exhibit DIF (Gierl et al., 1999; Shepard, Camilli, & Averill, 1981).

Although broad agreement exists on the definition of DIF, less agreement exists for which methods best detect DIF. These methods include the Mantel-Haenszel (MH), SIBTEST, standardization (STD), logistic regression, and item response theory (IRT) methods. A judgmental review approach, described by Gierl et al. (1999) and Holland and Thayer (1993), is sometimes used to identify the sources of DIF.

DIF can be thought of as differences in relative item difficulty that exaggerate or distort the actual group differences in ability (Camilli & Shepard, 1994). In IRT, the item parameters and the item characteristic curve for each item are assumed to be invariant across sub-populations. This property of invariance in item characteristic functions is tested in studies of DIF using IRT. As Thissen, Steinberg, and Wainer (1993) expressed it, the question to be answered is whether the estimated parameters for individual items differ significantly between the *focal* group and the *reference* group. According to Holland and Thayer (1988), in differential item functioning analysis, the group whose performance is of primary interest is the focal group, while the performance of the reference group is the standard against which the performance of the focal group is compared.

Translation and Curriculum as Sources of DIF

It has been well documented that even when rigorous processes of translation, verification, and field-testing are followed, translation may introduce measurement nonequivalence (Allalouf, 2000; Price & Oshima, 1998; Sireci & Swaminathan, 1996). Allalouf (2000), Allalouf, Hambleton, and Sireci (1999), and Gierl and Khaliq (2001), among others, have demonstrated approaches to identifying sources of translation DIF.

As Sireci and Swaminathan (1996) observed, the need to distinguish the effects of item language differences from those of language group differences complicates analyses of translated tests. For example, curriculum differences

such as the sequence of mathematics courses or the time spent on topics may cause DIF. Differential availability of textbooks and other materials for the language groups may also compound such differences.

Beyond differences in curriculum, the match between the curriculum and the content of the test is important. Other testing programs have examined this match. For example, Harnish and Linn (1981), Lawson, Bordignon, and Nagy (2002), Leinhardt and Seewald (1981), Mehrens and Phillips (1986), and Muthén, Kao, and Burstein (1991) investigated the effects of differences in instructional experiences of students on the resulting achievement estimates, latent trait definitions, and observed item difficulties. Indeed, several studies (e.g., Mehrens & Phillips, 1986; Miller & Linn, 1988) have suggested that the degree of match between an assessment and the curriculum can have a large impact on achievement test scores.

This Study

The data from Ontario students responding to the content subtest of the 2001 SAIP Mathematics Assessment provided a unique opportunity to investigate possible causes of DIF because Ontario has English- and French-language school boards. Not only did the language of instruction differ between these schools, but the curricula and available textbooks were also different. In addition, both 13- and 16-year-old students participated in the test and a new Ontario curriculum was introduced during the three years since the 16-year-old students were 13 years old. This study, therefore, addressed the following questions:

1. Do any items exhibit DIF between the English-language and French-language versions?
2. Can translation effects explain the DIF exhibited by any of these items? Can it be related to differences between the English-language and French-language curricula?

METHOD

The SAIP Mathematics Assessment

The Council of Ministers of Education, Canada (CMEC) develops and administers SAIP Mathematics, Science, and Language assessments to 13- and 16-year-old students across Canada in three- or four-year cycles. According to Fournier (2000), the SAIP provides insight into the factors affecting students' performances to determine whether the students in

different educational jurisdictions across Canada attain similar levels of performance at about the same age.

CMEC has administered the SAIP Mathematics Assessment three times: in 1993, 1997, and 2001. The data from the content subtest of the 2001 SAIP Mathematics Assessment were used for this study. The content subtest assesses student achievement in (a) Numbers and Operations, (b) Algebra and Functions, (c) Measurement and Geometry, and (d) Data Management and Statistics, and consists of 125 items comprising 75 multiple-choice items with four response options and 50 short-answer items. Both types of items were scored dichotomously. Students who wrote the content subtest received 27 background questions in addition to the 125 content items, which were administered in two stages. With the exception of the first 15 multiple-choice items, the remaining content items were organized by difficulty. Students were first administered a placement test, which consisted of the first 15 multiple-choice items in the full test. An exam proctor immediately scored these items. Based on the score on the placement test, each student was told to continue the test from one of three "starting points" — Item 16, Item 41, or Item 66 — and to work as far as possible within the test time (CMEC, 2001). To permit detailed analyses, this study focused on the 31 Measurement and Geometry items.

Sample

For the administration of the 2001 SAIP Mathematics Assessment, students were sampled from each participating province and, within some provinces, by language of instruction. To facilitate the examination of DIF caused by curriculum and language differences, we found it important to define two groups for comparison: the Ontario students in English-language and in French-language schools.

Of the Ontario students who took the test, those who omitted all 15 placement test items or did not provide their age were excluded. The resulting data set consisted of 793 13-year-olds and 677 16-year-olds from English-language schools and 487 13-year-olds and 546 16-year-olds from French-language schools.

For the analyses, students were divided by age and language. Missing items after the student's last response were treated as though the items had not been administered. For missing items before the student's last response, we assumed that the student read the item and chose not to respond. Because the placement test determined each student's starting point in the rest of the test, students responded to different subsets of the test items. The test developers expected that students would complete the 15 placement test

items plus 60 items from their assigned starting point. For this study, students' responses to Measurement and Geometry items on the placement test and within 60 items of their starting points were analyzed.

Predictions of Curricular and Translation Differences

In 1999, Ontario introduced a new high-school curriculum, which that year's grade-9 students used; this curriculum did not apply to earlier cohorts of students in grade 10 or higher in 1999. According to Ontario's "context statement," at the time of the 2001 SAIP Mathematics Assessment,

> most 13-year-old students were enrolled in either grade-8 or grade-9 mathematics, both of which are mandatory core subjects in the new curriculum . . . [however] most of the 16-year-old students in the assessment would have studied the old mathematics curriculum and taken a grade 11 course at one of the three possible levels of difficulty or would have taken no mathematics course since grade 10. (CMEC, 2001, p. 57)

The new curriculum differed from the old both in its content and in how it was developed. Before 1997, the provincial mathematics curriculum in Ontario was developed in English, and then translated into French (Ontario Ministry of Education, 1985a, 1985b). As a result, the defined curricula contained the same content in both languages, although differences existed in how the content was presented in textbooks and other resource materials, and different resources were available in English and in French. In contrast, the post-1997 mathematics curricula (Ontario Ministry of Education, 1997a, 1997b, 1999a, 1999b, 2000a, 2000b) were developed separately for French-language and English-language schools. The curriculum development teams worked in parallel and developed similar expectations for most of the content. However, a few expectations differed.

We had two important documents available to support our comparison of the curricula. The first, "SAIP 2001 Alignment with Ontario's Mathematics Curriculum," commissioned by the Ontario Ministry of Education, reported an analysis of each item on the 2001 SAIP Mathematics Assessment, indicating whether 13-year-old and 16-year-old students would have encountered it in the mathematics curricula. This analysis identified several items that addressed content that students would not have covered. For example, the SAIP assessment included several questions that could be solved using the sine or cosine formulas; neither 13- nor 16-year-old students would have been taught these formulas before taking the test. The document did not compare the English- and French-language curricula, however.

The second document, prepared by Ontario's Education Quality and Accountability Office (EQAO), was "Grade 9 Mathematics Curriculum

Ontario 1999 Expectation Mapping Chart for Developers." EQAO is responsible for developing and administering mathematics assessments based on Ontario's grade-9 French and English mathematics curricula. To support development of items for the assessments, EQAO staff created a document comparing the grade-9 English- and French-language curricula. This document identified the differences between the curricula of the English- and French-language schools.

In addition, three fluently bilingual mathematics educators reviewed the geometry items both for differences in difficulty of wording and curriculum differences.

IRT Calibration

We used BILOG-MG software (Zimowski, Muraki, Mislevy, & Bock, 1996), which performs multiple-group item response theory analysis for dichotomously-scored items, to compute two-parameter logistic (2PL) model maximum likelihood item parameter estimates for the 31 Measurement and Geometry items. The use of an IRT model allowed us to model both the relative difficulty of each item and its ability to distinguish among students with different levels of knowledge. The 2PL model was used because it provides clearer indications of DIF than do more complex models. The IRT calibrations were performed separately for the focal group (in this case, the students taking the French-language version of the test) and for the reference group (the students taking the English-language version). In the IRT calibration and scoring analyses, omitted items were counted as incorrect, and "not presented" items (i.e., items skipped because of placement test assignment or beyond the last mark on the answer sheet) were not included in the analyses. We used BILOG-MG to compute examinee score estimates, based on the obtained item parameter estimates. Expected *a posteriori* (EAP) scoring was used, so that a $N(0,1)$ population prior was incorporated into the estimates.

DIF Analyses

We used LINKDIF (Waller, 1998) to compute DIF indices. LINKDIF linked the separate item parameter estimates to a common metric using the test characteristic curve method of Stocking and Lord (Stocking & Lord, 1983). The Stocking and Lord procedure is a characteristic curve equating method that estimates linking coefficients (to be applied to the a and b parameters in the secondary calibration) by minimizing the difference between the original test characteristic curve (TCC) and that based on the transformation.

Once the item parameters were linked, LINKDIF computed the Lord's χ^2 (Lord, 1980) and associated significance levels.

Lord's chi-square statistic is an index of DIF. It is computed as

$$\chi_i^2 = \mathbf{v}_i' \mathbf{S}^{-1} \mathbf{v}_i$$

where

$\mathbf{v}_i'$ is a vector of the differences in the estimated item parameters for the ith item between the focal and reference groups, and

$\mathbf{S}_i$ is the asymptotic variance-covariance matrix for the differences in item parameter estimates (Lord, 1980).

It follows a chi-square distribution, so that values of the index can be compared to the critical value corresponding to a specified alpha level with 3 degrees of freedom under the null hypothesis of no DIF. Because tests were performed for 31 items, we used an alpha level equal to .0016 (.05/31).

RESULTS AND DISCUSSION

Predictions of Curricular and Translation Differences

Curricula. The second and third columns of Table 1 summarize the Ontario Ministry of Education's analysis of the extent to which the knowledge and skills required to answer each of the SAIP items were taught in the old curriculum (for 16-year-old students) and the new curriculum (for 13-year-old students). These columns predict which items may be difficult both for students taking the French-language version and those taking the English-language version. A similar item-by-item analysis of curriculum-based differences in difficulty *between* the French- and English-language versions provided only a few specific predictions because, as described above, the differences are not in the content specified, but in the level of detail with which it is specified. For example, in *Analytical Geometry*, the curriculum for French-language, grade-9 applied courses requires students to "communiquer et justifier, de façon claire et concise, les étapes de son raisonment dans le dévelopment d'une solution" and to "utiliser la terminologie et la notation appropriée au plan cartésien" (Ontario Ministry of Education, 1997a, p. 20), while that of English language schools simply requires students to "communicate solutions in established mathematical form, with clear reasons given for the steps taken" (Ontario Ministry of Education, 1997b, p. 23). As this example illustrates, the differences are generally not in the content to be taught, but rather in the level of detail with which it is described.

TABLE 1

Measurement and Geometry Items on the Content Subtest of the 2001 SAIP Mathematics Assessment: Predicted Curricular and Translation Differences

	SAIP to Ontario Curriculum Alignment[1]		Grade 9 Expectation Mapping[2]		Review for Translation and Taught Curriculum Differences[3]	
Item	13-Year-Olds	16-Year-Olds	Comments	Prediction	Comments	Prediction
1	No	Yes				
10	Yes	Yes				
12	Not quite	Not likely				
15	Not quite	Yes				
23	Yes	Yes			Item refers to pencils. Teachers in French-language schools tend to encourage students to use pencils.	Easier in French
24	Yes	Not quite				
26	Yes	Yes				
33	Yes	Yes			The use of the word "lignes" in French to mean both straight and curved lines may be confusing.	Easier in English
35	Yes	Yes			Item includes a map of Quebec City. Students in French-language schools are more likely to be familiar with Quebec City.	Easier in French
36	Yes	Yes				
42	Yes	Yes				
47	Yes	Yes			Item includes a drawing of a triangular prism. Students may be more used to seeing prisms drawn as "nets" than as solid objects.	Difficult in both
49	Not quite	Yes			Item requires knowledge of the terms *edges*, *faces*, and *vertices*. French-language schools tend to place more emphasis on terminology.	Easier in French
53	Yes	Yes			Item mixes centimeters and meters. French-language schools tend to drill more on units.	Easier in French
64	Yes	Yes				
65	Yes	Yes				

TABLE 1, CONTINUED

	SAIP to Ontario Curriculum Alignment[1]		Grade 9 Expectation Mapping[2]		Review for Translation and Taught Curriculum Differences[3]	
Item	13-Year-Olds	16-Year-Olds	Comments	Prediction	Comments	Prediction
69	Yes	Yes				
74	Yes	Yes				
83	Yes	Yes				
84	No	Yes				
86	No	Maybe				
88	No	No				
94	No	Maybe			Item requires more reading than most items and the vocabulary in the French version is more difficult than in the English.	Easier in English
96	Yes	Yes	Can be solved using the Pythagorean theorem	Easier in French		
100	No	Not likely	Can be solved using the Pythagorean theorem	Easier in French		
101	No	No				
105	No	No				
108	Not quite	Yes	Can be solved using the Pythagorean theorem	Easier in French	Item requires an application of the Pythagorean theorem to variables *d, w, x, y,* and *z*. Students may not have encountered the square-root sign with variables before.	Difficult in both
109	Yes	Yes				
110	No	Yes			Item requires drawing a diagram, then performing a difficult calculation.	Difficult in both
125	No	No	Can be solved using the Pythagorean theorem	Easier in French		

Notes

1 Based on an analysis commissioned by the Ontario Ministry of Education.

2 Based on an analysis by Ontario's Education Quality and Accountability Office.

3 Based on a review by three bilingual mathematics teachers.

The fourth and fifth columns in Table 1 summarize relevant predictions based on the comparison of the grade-9 mathematics curricula. Although the Pythagorean theorem is taught to all students, it is mentioned explicitly in both the overall and specific expectations for the French-language curriculum for grade 9, but nowhere in the English-language curriculum for grade 9. It is possible that the items that can be solved using the Pythagorean theorem may be easier for students taking the French-language version because they may have had more instruction on that content.

Columns six and seven summarize the review of the items by the bilingual mathematics educators. They identified some differences in classroom practice unrelated to the curriculum. For example, they suggested that teachers in French-language schools spend more time drilling students on the use of measurement units. Because Item 53 mixes centimetres and metres, they predicted that students taking the French-language version might find it slightly easier. In addition, they noted that differences in what students know are not always related to the curriculum. Item 35, for example, includes a map of the Quebec City area, which is likely to be more familiar to students in French-language schools than to students in English-language schools.

Translation. Each item on the SAIP tests was developed in either English or French and then translated into the other language. As the report of the 1997 mathematics assessment (CMEC, 1997) describes,

> A linguistic analysis of each question and problem was also conducted to make sure French and English items functioned in the same manner. For the marking sessions, francophone and anglophone coders were jointly trained and did the marking together in teams working in the same rooms. (p. 4)

We would expect these efforts to minimize the possible sources of translation DIF. As other studies (e.g., Allalouf et al., 1999; Gierl & Khaliq, 2001) have shown, however, it is very difficult to achieve perfect agreement in the meaning and vocabulary difficulty of translated materials.

Beyond vocabulary difficulty, more subtle translation differences may occur. As summarized in columns six and seven of Table 1, the teachers' reviews of the items suggested some possible differences. For example, Item 49 involves edges, faces, and vertices of a three-dimensional object. The teachers reported that, in their experiences, the French-language schools emphasize knowledge of terminology more than do the English-language schools. In addition, the term "side" is often used instead of "face" in the materials for the English-language students, which may make the item particularly confusing for English-language students.

Items Exhibiting DIF

Table 1 also summarizes predicted curriculum and translation differences between 13- and 16-year-old students and between English- and French-language students. It shows that the bilingual educators predicted that eight of the Measurement and Geometry items would be easier for students taking the French-language version of the test while two items would be easier for students taking the English-language version. We found that 7 of the 10 items predicted to have differential difficulty by the bilingual educators or the curriculum comparison show statistically significant DIF in the direction predicted.

The DIF analyses flagged seventeen of the thirty-one Measurement and Geometry items as exhibiting DIF for the 13-year-olds, while five were flagged as exhibiting DIF for the 16-year-olds. Table 2 presents these results. Thirteen of these items are the last thirteen Measurement and Geometry items. All but two of the items flagged were easier for students taking the French-language version of the test. Four of the five items exhibiting DIF for 16-year-old students are the last four Measurement and Geometry items. All five items are easier for students taking the French-language version of the test. Of the seventeen items that showed DIF for the 13-year-olds, eleven are multiple-choice items while six are short-answer items. Four of the five DIF items for the 16-year-old students are multiple-choice items. Only four items were flagged for both 13- and 16-year-old students.

Table 3 summarizes the differences predicted by the bilingual teachers and the curriculum mapping and those found based on the DIF analyses. As this table illustrates, some correspondence occurred between the predictions and the DIF analysis results. Ten items were predicted by the bilingual educators or the comparisons of the English- and French-language curricula to exhibit DIF; seven of these (70%) exhibited DIF in the predicted direction. Eighteen items exhibited statistically significant DIF for 13-year-olds, 16-year-olds, or both; only 7 (39%) were predicted to exhibit DIF in the found direction.

A Plausible Alternative Explanation for DIF

Examination of the positions in the test of the items exhibiting DIF suggested an alternative explanation. As Figure 1 illustrates, 13-year-old students taking the English-language version of the test attempted more items than did students taking the French-language version. The pattern is also similar for 16-year-old students. This makes the interpretation of the DIF analysis results difficult. Although more students taking the English-language

TABLE 2

Measurement and Geometry Items on the Content Subtest of the 2001 SAIP Mathematics Assessment: DIF Analysis Results

Item Order	Item Type	Achievement Level	Target Ability	— 13-Year-Old Students — English *a*	English *b*	French *a*	French *b*	Lord's c^2 Index	— 16-Year-Old Students — English *a*	English *b*	French *a*	French *b*	Lord's c^2 Index
1	MC	3	C	0.71	–1.03	0.51	–1.82	7.09	0.85	–2.15	0.78	–2.28	0.18
10	MC	3	PS	0.84	–0.14	0.79	–0.04	1.28	0.59	–0.93	0.78	–0.78	3.08
12	MC	3	C	0.51	0.48	0.58	0.29	1.69	0.85	–0.25	0.86	–0.12	2.01
15	MC	3	C	0.48	0.48	0.62	0.22	4.31	0.74	–0.38	1.16	–0.44	12.39
23	MC	1	C	0.48	–2.76	0.03	–4.73	33.24*	0.74	–1.92	0.56	–2.47	2.08
24	MC	1	C	0.03	–9.60	0.03	–4.95	0.32	0.65	–2.40	0.56	–2.65	0.40
26	MC	1	C	0.03	–9.57	0.03	–4.73	0.33	0.67	–3.05	0.77	–2.73	0.30
33	SA	1	C	0.51	–3.28	0.04	–4.37	24.84*	0.99	–2.22	1.02	–1.88	5.45
35	SA	1	P	0.59	–1.62	0.05	–3.95	59.17*	1.02	–1.60	0.93	–1.65	0.36
36	SA	1	P	0.35	–2.43	0.29	–2.33	4.36	0.44	–2.04	0.79	–1.71	10.43
42	MC	2	C	0.48	–2.81	0.04	–6.29	30.00*	0.86	–1.87	0.97	–1.76	0.43
47	MC	2	PS	0.54	–0.91	0.64	–0.72	1.67	0.82	–1.14	1.39	–1.02	14.18
49	MC	2	C	0.49	–1.34	0.46	–1.08	3.72	0.64	–1.09	0.72	–1.42	11.79
53	MC	2	PS	0.74	–0.28	0.61	–0.24	2.02	0.85	–0.67	0.91	–0.96	11.24
64	SA	2	P	0.89	–0.67	0.93	–0.65	0.13	1.10	–1.04	0.94	–1.16	1.48
65	SA	2	PS	0.38	1.14	0.35	0.99	1.66	0.63	–0.20	0.89	–0.21	5.77
69	SA	3	PS	0.55	–0.71	0.59	–0.89	2.79	0.82	–0.84	1.08	–1.06	16.12*
74	SA	3	C	0.77	0.69	0.90	0.48	3.65	0.83	–0.14	1.22	–0.25	10.19
83	MC	4	C	0.80	1.82	0.44	1.46	43.82*	0.96	0.54	1.44	0.33	11.75
84	MC	4	PS	0.47	1.90	0.24	0.86	57.46*	0.62	0.55	0.74	0.32	3.36
86	MC	4	C	0.49	3.90	0.12	4.27	20.86*	1.09	1.61	0.81	1.95	2.72
88	MC	4	C	0.58	2.05	0.26	1.56	51.65*	0.90	1.27	0.79	1.13	4.92
94	SA	4	P	0.70	3.14	0.20	3.51	27.11*	0.87	1.90	1.10	1.79	2.21
96	SA	4	P	0.79	2.13	0.17	2.00	59.50*	1.08	0.90	1.01	0.81	2.00
100	SA	4	PS	0.28	6.13	0.15	3.05	126.97*	1.14	1.65	0.80	1.88	5.07
101	MC	5	C	0.29	1.20	0.20	–0.43	20.96*	0.66	0.40	0.57	0.65	2.36
105	MC	5	P	0.32	2.12	0.15	0.48	40.19*	1.04	1.01	0.72	0.87	16.39*
108	MC	5	C	0.34	2.18	0.18	–0.11	46.01*	1.21	1.22	0.63	0.78	59.09*
109	MC	5	PS	0.43	2.91	0.18	–0.04	87.68*	1.18	1.42	1.10	0.98	23.96*
110	MC	5	PS	0.27	2.28	0.17	0.11	32.71*	0.73	1.70	0.63	1.21	21.65*
125	SA	5	PS	0.29	2.32	0.19	–0.15	39.34*	0.47	1.69	0.47	1.48	1.21

Notes:

MC = multiple-choice; SA = short-answer; C = conceptual; P = procedural; PS = problem solving; * $p < .0016$. The *a* and *b* parameters for the English and French versions of the test have been equated. Item classifications were provided by the Council of Ministers of Education, Canada.

TABLE 3

Measurement and Geometry Items on the Content Subtest of the 2001 SAIP Mathematics Assessment: Comparison of Predictions and Statistical Analysis Results

Item Order	Grade 9 Bilingual Educators' Prediction: Easier in…	Expectation Mapping Prediction Easier in…	DIF — Statistics — 13-Year-Olds	DIF — Statistics — 16-Year-Olds	Comments
1					Even though 13-year-old students had not been taught this content, they performed well.
10					
12					Even though this content is not in Ontario's old or new curriculum, both 13- and 16-year-old students performed well.
15					Even though 13-year-old students had not been taught this content, they performed well.
23	French		French		Very low *a* parameter for 13-year-old students in French-language schools
24					Very low *a* parameters for 13-year-old students in French- and English-language schools
26					Very low *a* parameters for 13-year-old students in French- and English-language schools
33	English		French		Very low *a* parameter for 13-year-old students in French-language schools
35	French		French		Very low *a* parameter for 13-year-old students in French-language schools
36					
42			French		Very low *a* parameter for 13-year-old students in French-language schools
47					
49	French				
53	French				
64					
65					
69				French	
74					
83			French		
84			French		
86			English		
88			French		
94	English		English		
96		French	French		
100		French	French		This item was very difficult for all examinees.
101			French		
105			French	French	
108		French	French	French	
109			French	French	
110			French	French	
125		French	French		

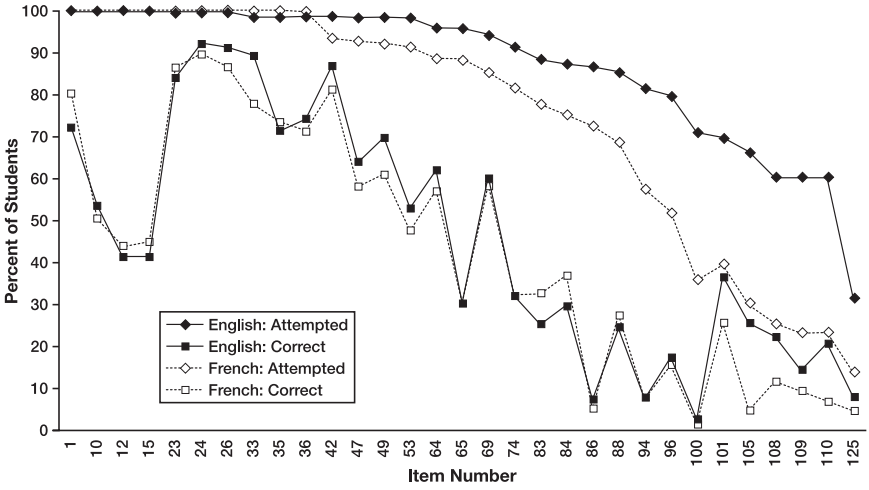

Figure 1. Percentages of English- and French-language 13-year-old students assigned to answer each item who attempted the item and who answered it correctly.

version attempted the later items and also tended to answer more of these items correctly, the percentages of students answering each item correctly out of those who attempted it are generally higher for students taking the French-language version, indicating that they were more likely, if they responded to an item, to respond correctly. Different explanations are possible. It may be that the Ontario students taking the French-language version found the vocabulary used difficult which in turn resulted in slower responding. However, it is also possible that other factors, such as experience with similar tests or a lesser propensity to guess, contributed to a different test-taking approach. These possibilities merit further investigation.

Limitations and Future Directions

These results must be interpreted with caution because the need to limit the analyses to students in one province so as to permit close comparisons of the curricula by languages resulted in small sample sizes. The sizable proportions of students not attempting items increased the difficulty of fitting the models. The model fit was particularly problematic for the 13-year-old students taking the French-language version.

This study has other limitations. The correspondence between the prescribed curriculum and what is actually taught in the classroom is rarely perfect, complicating the prediction of which items are likely to exhibit DIF. Although few differences were found between mathematics curricula for the two languages of instruction in Ontario, it is not easy to dismiss curriculum as a possible source of DIF without analyzing actual classroom experiences of the two groups. A survey of teachers in both French- and English-language schools regarding their perceptions of the difficulty of the items and of the differences in the taught curriculum might provide additional information about what items might function differentially.

The results suggest areas for further exploration. Curricular differences might be better understood in combination with information about teachers' classroom practices. Teachers' academic training, experience, and the materials available to them might well influence practices. Such contextual information would help us understand how the curriculum is being understood and presented. In addition, an examination of the patterns of items attempted by students taking the English- and French-language versions suggests a difference in test-taking approaches. Further research on the test-taking approaches of these two groups might well explain some of the differences in test results.

CONCLUSION

The purpose of this study was to investigate the possible impacts not only of language, but also of curriculum differences, on how students from different subpopulations performed on test items. By focusing on Measurement and Geometry items and students in Ontario who took French- and English-language versions of the test, we were able to explore such differences more closely than would have been possible had we used more diverse sets of items and samples of students. The results illustrate the complexity of the factors that contribute to how items are understood and answered by different groups of students. Although the results are not conclusive, this study demonstrates the importance of such analyses and, we hope, will provide a starting point for future studies.

ACKNOWLEDGEMENTS

The authors would like to thank the Council of Ministers of Education, Canada for providing the data, and the Ontario Ministry of Education and the Education Quality and Accountability Office for the use of their documentation. Also, thanks are due to Alana Cote, Hervé Jodouin, Kelsey Saunders, and Jacques Theoret of Ontario's Education Quality and Accountability Office and to Gilles Fournier of the CMEC

for their helpful suggestions throughout this study.

Correspondence concerning this article should be addressed to Ruth A. Childs, OISE/UT, 252 Bloor Street West, 11th Floor, Toronto, Ontario M5S 1V6. E-mail: rchilds@oise.utoronto.ca.

REFERENCES

Allalouf, A. (2000, April). *Retaining translated verbal reasoning items by revising DIF items.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Boiteau, N., Bertrand, R., & St-Onge, C. (2002, May). *Evolution of French-English DIF items in SAIP math data.* Paper presented at the annual meeting of the Canadian Society for Studies in Education, Toronto.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Council of Ministers of Education, Canada. (1997). *School Achievement Indicators Program (SAIP) 1997 Mathematics II Assessment.* Toronto: Author.

Council of Ministers of Education, Canada. (2001). *Report on Mathematics Assessment III. School Achievement Indicators Program (SAIP) 2001.* Toronto: Author.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2002, May). *Comparability of English and French versions of SAIP for reading, mathematics, and science items.* Paper presented at the annual meeting of the Canadian Society for Studies in Education, Toronto.

Fournier, G. (2000). The pan-Canadian assessments: Setting the records straight. *Phi Delta Kappan, 81*, 547–550.

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

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

Harnish, D., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133–146.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Thayer, D. T. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Koh, K., & Ercikan, K. (2002, May). *Construct comparability in French and English versions of SAIP.* Paper presented at the annual meeting of the Canadian Society for Studies in Education, Toronto.

Lawson, A., Bordignon, C., & Nagy, P. (2002). Matching the Grade 8 TIMSS item pool to the Ontario curriculum. *Studies in Educational Evaluation, 28,* 87–102.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18*, 85–95.

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

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*, 185–196.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*, 205–219.

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Applications of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1–22.

Ontario Ministry of Education. (1985a). *Curriculum guideline, mathematics intermediate and senior divisions: Part two, Grades 7and 8, Grades 9 and 10, general level, Grades 11 and 12, general level.* Toronto, ON: Author.

Ontario Ministry of Education. (1985b). *Programme-cadre mathématiques, cycles intermédiaire et supérieur: Deuxième partie, 7e et 8e année, 9e et 10e année, niveau général, 11e et 12e année, niveau général.* Toronto: Author.

Ontario Ministry of Education. (1997a). *Le curriculum de l'Ontario, de la 1re la 8e année — Mathématiques.* Toronto: Author.

Ontario Ministry of Education. (1997b). *The Ontario curriculum, Grades 1-8 — Mathematics.* Toronto: Author.

Ontario Ministry of Education. (1999a). *Le curriculum de l'Ontario, 9e et 10e année — Mathématiques.* Toronto: Author.

Ontario Ministry of Ed ucation. (1999b). *The Ontario curriculum, Grades 9 and 10 — Mathematics.* Toronto: Author.

Ontario Ministry of Education. (2000a). *Le curriculum de l'Ontario, 11e et 12e année — Mathématiques.* Toronto: Author.

Ontario Ministry of Education. (2000b). *The Ontario curriculum, Grades 11 and 12, Mathematics.* Toronto: Author.

Price, L. R., & Oshima, T. C. (1998, April). *Differential item functioning and language translation: A cross-national study with a test developed for certification.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Shepard, L., Camilli, G., & Averill, A. (1981) Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*(4), 317–375.

Sireci, S. G., & Swaminathan, H. (1996, October). *Evaluating translation equivalence: So what's the big dif?* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*(1), 53–66.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*(2), 89–99.

Waller, N. G. (1998). LINKDIF: Linking item parameters and calculating IRT measures of differential fuctioning of items and test [Computer Program Exchange]. *Applied Psychological Measurement, 22,* 392.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multi-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*(2), 225–247.