

Running head: SOURCES OF TRANSLATION DIF

**Identifying Sources of Differential Item Functioning on Translated
Achievement Tests: A Confirmatory Analysis****

Mark J. Gierl

Shameem Nyla Khaliq

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Paper Presented at the Annual Meeting of the
National Council on Measurement in Education (NCME)

New Orleans, Louisiana, USA

April 24-27, 2000

** This paper can also be downloaded from the Centre for Research in Applied Measurement and Evaluation (CRAME) website: <http://www.education.ualberta.ca/educ/psych/crame/>

Abstract

Increasingly, tests are being translated and adapted into different languages. Differential item functioning (DIF) analyses are often used to identify non-equivalent items across language groups. However, few studies have focused on understanding why some translated items produce DIF. The purpose of the current study is to identify sources of differential item functioning on translated achievement tests using substantive and statistical analyses. A substantive analysis of existing DIF items was conducted by an 11-member committee of test developers, analysts, editors, and translators. In their review, four sources of translation DIF were identified. Two certified translators used these four sources to categorize a new set of DIF items from Grade 6 and 9 Mathematics and Social Studies Achievement Tests. Each bundle was associated with a specific source of translation DIF and each bundle was anticipated to favor a specific group of examinees. Then, a statistical analysis was conducted on each bundle using SIBTEST. The translators categorized the mathematics DIF items into three sources, and they correctly predicted the group that would be favored for seven of the eight bundles across the two grade levels. The translators categorized the social studies DIF items into four sources, and they correctly predicted the group that would be favored for eight of the 13 bundles across the two grade levels. The majority of items in mathematics and social studies were associated with differences in the words, expressions, or sentence structure of items that are not inherent to the language and/or culture. By combining substantive and statistical DIF analyses, researchers can study the sources of DIF and create a body of confirmed DIF hypotheses that may be used to develop guidelines and test construction principles for reducing DIF on translated tests.

Identifying Sources of Differential Item Functioning on Translated Achievement Tests: A Confirmatory Analysis

Increasingly, educational and psychological tests are being translated and adapted for use in different languages and cultures. Many examples can be found. At the international testing level, the International Association for the Evaluation of Educational Achievement (IEA) conducted the Third International Mathematics and Science Study in 1995, and administered tests in 31 different languages to 45 participating countries. At the national testing level, the Council of Ministers of Education in Canada assesses achievement of 13- and 16-year-old students in reading and writing, mathematics, and science in English and French for the provinces and territories. At the local testing level, the Department of Learning in the Canadian province of Alberta translates eight of their 11 high school exiting exams into French to fulfill the national language requirement. These trends are expected to continue. The IEA is conducting a fourth study, and major efforts have been undertaken to develop an equivalent set of tests in numerous languages. Hambleton (1994) speculates that test adaptations and translations will become even more prevalent in the future due to an increase in international testing, more demand for credentialing and licensure exams in multiple languages, and a growing interest in cross-cultural research (for some recent examples, see Hambleton & Patsula, 1998; Jeanrie & Bertrand, 1999; Reckase & Kunce, 1999).

Differential item functioning (DIF) analyses are often used during the test translation and adaptation process to identify items that function differently between language groups. DIF is present when examinees from different groups have a different probability or likelihood of answering an item correctly, after conditioning on overall ability (Shepard, Camilli, & Averill, 1981). Researchers who study the psychometric characteristics of translated tests have noted an important trend: The amount of DIF on some translated tests is large. For example, Gierl, Rogers, and Klinger (1999) reported that 26 of 50 items (52%) on a Canadian Grade 6 social studies achievement test translated from English to French displayed moderate or large DIF. Ercikan (1999) found that 58 out of 140 science items (41%) on the Third International Mathematics and Science Study (TIMSS) displayed moderate or large DIF when the Canadian English and French examinees were compared. She also found that 29 out of the 158

mathematics items (18%) displayed DIF. Allalouf, Hambleton, and Sireci (1999) noted that 42 out of 125 verbal items (34%) displayed moderate or large DIF on the Israeli Psychometric Entrance Test when Hebrew and Russian examinees were compared. These outcomes raise questions and concerns about the validity of tests that are translated or adapted from one language group to another. These findings also highlight the need to identify the sources and causes of translation DIF so that steps can be taken during test development to improve the exams.

Surprisingly, few studies have focused on why translated test items function differently across languages (Allalouf et al., 1999; Hambleton, 1994). And yet the implications of this research are clear: If the sources of translation DIF could be anticipated, then test developers could carefully monitor their test construction, translation, and adaptation practices to ensure the different language forms of the exam are comparable across language groups. Research focusing on the sources of translation DIF also could lead to new policies and practices (e.g., item writing guidelines) that might reduce the number of items that do not function equivalently across languages.

Although the number of studies is limited, researchers are beginning to systematically identify the sources of DIF on translated tests. For example, Allalouf et al. (1999) studied translated verbal items that functioned differently for Hebrew- and Russian-speaking examinees on the Israeli Psychometric Entrance Test in order to identify the possible causes. Allalouf et al. had five Hebrew-to-Russian translators complete a blind item review of 60 items. For the review, the translators were asked to predict specific item characteristics including whether the item displayed DIF, the direction of DIF, the magnitude of DIF, and the reason for DIF. After the reviews were completed, an 8-member committee (the original five translators and three Hebrew-speaking researchers) reviewed the DIF items, evaluated the suggested causes of DIF, and reached agreement on these causes for each DIF item. The 8-member committee identified four probable causes of translation DIF: (a) differences in cultural relevance, (b) changes in difficulty of words or sentences, (c) changes in format, and (d) changes in content. Allalouf et al. note that their results may be limited because only two language groups were used. Replication, they

argue, would help to validate the reasons for DIF and to ensure the findings are generalizable across language groups.

The purpose of the current study is to identify sources of differential item functioning on translated achievement tests. This study sheds light on the substantive hypotheses put forward by Allalouf et al. (1999) by identifying the sources of DIF on translated achievement tests using different language groups, content areas, and grade levels. It is based on the approach described by Roussos and Stout (1996a) where substantive and statistical analyses are used to identify and evaluate DIF hypotheses. By combining substantive and statistical DIF analyses, researchers can study the sources of DIF and create a body of confirmed DIF hypotheses that, in turn, can be used to develop guidelines and test construction principles for reducing DIF on translated tests.

Multidimensional DIF Analysis Paradigm

Roussos and Stout (1996a) proposed a paradigm that unifies substantive and statistical analyses by linking both to the Shealy-Stout multidimensional model for DIF (MMD, Shealy & Stout, 1993). The need for an alternative approach was motivated by the inability of content specialists to identify the sources of DIF consistently and accurately. Typically, DIF statistical analyses are followed by content reviews to identify sources of DIF. Reviewers are asked to study DIF items and identify why these items are more difficult for one group of examinees compared to another (see, for example, Berk, 1982; Ramsey, 1993). But researchers found that reviewers were generally poor at predicting which items would function differently across groups (Englehard, Hansche, & Rutledge, 1990; Gierl & McEwen, 1998; Plake, 1980; Rengel, 1986; Sandoval & Miille, 1980). Experience has also demonstrated that it is difficult to interpret DIF using a judgmental approach (Angoff, 1993; Bond, 1993; Camilli & Shepard, 1994; O'Neill & McPeck, 1993). Based on a review of the DIF literature, Roussos and Stout (1996a) concluded: "Attempts at understanding the underlying causes of DIF using substantive analyses of statistically identified DIF items have, with few exceptions, met with overwhelming failure" (p. 360). To overcome this problem, they proposed a confirmatory approach conducted in two stages. The first stage is a substantive analysis where DIF hypotheses are generated. The

second stage is a statistical analysis of the DIF hypotheses. By combining substantive and statistical analyses in a confirmatory framework, Roussos and Stout contend, researchers can identify and study the causes of DIF. This approach also provides some Type I error control since only a small number of DIF hypotheses are tested (Stout & Roussos, 1995).

Shealy-Stout Multidimensional Model for DIF (MMD)

MMD is a framework for understanding how DIF occurs. It is based on the assumption that multidimensionality produces DIF. A dimension is a substantive characteristic of an item that can affect the probability of a correct response on the item. The main construct that the test is intended to measure is the primary dimension. DIF items measure at least one dimension in addition to the primary dimension (Ackerman, 1992; Roussos & Stout, 1996a; Shealy & Stout, 1993). The addition of dimensions that produce DIF are referred to as the secondary dimensions. When primary and secondary dimensions characterize item responses, the data are considered multidimensional. Secondary dimensions are considered auxiliary if they are assessed intentionally as part of the test construct. Alternatively, the secondary dimensions are considered nuisance if they are unintentionally assessed as part of the test construct. DIF that is caused by auxiliary dimensions is benign because the test is expected to measure these dimensions whereas DIF that is caused by nuisance dimensions is adverse because the test is not expected to measure these dimensions. When conducting DIF analyses, it is a matter of judgement as to whether the secondary dimension is interpreted as benign or adverse in a particular testing situation because the purpose of the test, the nature of the secondary dimension, and the examinees of interest must be considered.

Substantive DIF Analysis

The Roussos-Stout DIF analysis paradigm is built on the foundation provided by MMD. The first stage is a substantive analysis where DIF hypotheses are generated. The DIF hypothesis specifies whether an item or bundle (i.e., two or more items) designed to measure the primary dimension also measure a secondary dimension, thereby producing DIF, for examinees in either the reference or the focal group. The reference group is the majority group or the group to whom the focal group is compared. The focal group is the minority group or the particular group of

interest in the DIF analysis. Roussos and Stout (1996a) believe that MMD can be used to design DIF-free items if test developers can generate accurate DIF hypotheses based on their understanding of the underlying dimensional structure of the test. Roussos and Stout outline four methods that can help test developers specify this structure so DIF hypotheses can be formulated.

First, previous published DIF analyses can be used. In these studies, the primary and secondary dimensions are often outlined by the authors who provide interpretations of their DIF results. Second, substantive content considerations and judgements can help generate DIF hypotheses. Test developers are content specialists who can use their substantive knowledge of a content area to identify primary and secondary dimensions and to predict whether one group of examinees will outperform a second group of examinees due to the examinees' proficiency on these dimensions. Third, archival test data can be analyzed when specific secondary dimensions are prominent. For example, many large-scale testing programs develop content-specific tests to measure more general constructs such as verbal or mathematical reasoning. Moreover, Roussos and Stout (1996a, p. 355) note that many standardized tests require examinees to solve items in a meaningful context which, despite the obvious benefits, may also pose a risk for test equity and fairness because specific groups of examinees may be differentially familiar with test contexts thereby providing a unfair advantage for one group over another. By identifying these contexts and, subsequently, the possible secondary dimensions that are produced by these contexts from analyzing archival data, DIF hypotheses can be specified. Fourth, testing bundles of items according to some "organizing principle" rather than individual items can lead to dimensionality-based DIF hypotheses when the bundles reflect a secondary dimension (Douglas, Roussos, & Stout, 1996).

Statistical DIF Analysis

The second stage in the Roussos-Stout multidimensionality-based DIF paradigm is confirmatory statistical testing of the DIF hypotheses. The Simultaneous Item Bias Test (SIBTEST) can be used to statistically test DIF hypotheses and quantify the size of DIF. With this statistical approach, the complete latent space is viewed as multidimensional, (Θ, η) , where Θ is

the primary dimension and η is the secondary dimension. SIBTEST is designed to identify items or bundles in the secondary dimension. It can also be used to estimate the amount of DIF.

The statistical hypothesis tested by SIBTEST is:

$$H_0: \mathbf{b}_{UNI} = 0 \text{ vs. } H_1: \mathbf{b}_{UNI} \neq 0,$$

where \mathbf{b}_{UNI} is the parameter specifying the amount of DIF for an item or bundle. \mathbf{b}_{UNI} is defined as

$$\mathbf{b}_{UNI} = \int B(\Theta) f_F(\Theta) d\Theta,$$

where $B(\Theta) = P(\Theta, R) - P(\Theta, F)$, the difference in the probabilities of correct response for examinees from the reference and focal groups, respectively, conditional on Θ , $f_F(\Theta)$ is the density function for Θ in the focal group, and d is a scaling constant. \mathbf{b}_{UNI} is integrated over Θ to produce a weighted expected score difference between reference and focal group examinees of the same ability on an item or bundle. To operationalize this statistical DIF analysis, items on the test are divided into the studied subtest and the matching subtest. The studied subtest contains the item or bundle believed to measure the primary and secondary dimensions whereas the matching subtest contains the items believed to measure only the primary dimension. Matching subtest scores are used to place the reference and focal group examinees into subgroups at each score level so their performances on items from the studied subtest can be compared. More specifically, items 1...n denote the matching subtest items and items n+1...N denote the studied subtest items. In the case of a single-item DIF analysis, only one item is included in the studied subtest but, in a bundle analysis, two or more items are included in the studied subtest. U_i denotes the response to item i scored 0 or 1. For each

examinee, $X = \sum_{i=0}^n U_i$ to specify the total score on the matching subtest and $Y = \sum_{i=n+1}^N U_i$ to

specify the total score on the studied subtest. Examinees in the reference and focal groups are then grouped into K subgroups based on their matching subtest scores so that the examinees

from each group with the same matching subtest score can be compared. After the matching is complete, examinee performance on the studied subtest item or bundle can be assessed.

The weighted mean difference between the reference and focal groups on the studied subtest item or bundle across the K subgroups is given by

$$\widehat{\mathbf{b}}_{UNI} = \sum_{k=0}^K p_k d_k,$$

which provides an estimate of \mathbf{b}_{UNI} . In this equation, p_k is the proportion of focal group examinees in subgroup k and $d_k = \overline{Y_{Rk}^*} - \overline{Y_{Fk}^*}$, which is the difference in the adjusted means on the studied subtest item or bundle for the reference and focal groups, respectively, in each subgroup k. The means on the studied subtest item or bundle are adjusted to correct for any differences in the ability distributions of the reference and focal groups using a regression correction described in Shealy and Stout (1993; also see Jiang & Stout, 1998).

SIBTEST yields an overall statistical test for $\widehat{\mathbf{b}}_{UNI}$. In addition, $\widehat{\mathbf{b}}_{UNI}$ can be interpreted as the amount of DIF for each item or bundle. Positive values of $\widehat{\mathbf{b}}_{UNI}$ indicate DIF favoring the reference group and negative values indicate DIF favoring the focal group. The test statistic for testing the null hypothesis is

$$SIB = \frac{\widehat{\mathbf{b}}_{UNI}}{\widehat{\mathbf{s}}(\widehat{\mathbf{b}}_{UNI})}$$

where $\widehat{\mathbf{s}}(\widehat{\mathbf{b}}_{UNI})$ is given by

$$\widehat{\mathbf{s}}(\widehat{\mathbf{b}}_{UNI}) = \left[\sum_k p_k^2 \left(\frac{1}{N_{R_k}} \widehat{\mathbf{s}}^2(Y|k, R) + \frac{1}{N_{F_k}} \widehat{\mathbf{s}}^2(Y|k, F) \right) \right]^{1/2}.$$

\widehat{p}_k is the proportion of examinees in the focal group obtaining $X = k$ on the matching subtest, N_{R_k} and N_{F_k} are the sample sizes for the reference and focal examinees, respectively, with a

total score of k on the matching subtest, and $\widehat{S}^2(Y|k, R)$ and $\widehat{S}^2(Y|k, F)$ are the sample variances for reference and focal group examinees, respectively, on the studied subtest (which is either an item or bundle) with a total score of k on the matching subtest. Shealy and Stout (1993) demonstrated that SIB has a normal distribution with mean 0 and variance 1 under the null hypothesis of no DIF. The null hypothesis is rejected if SIB exceeds the 100 $(1 - \alpha / 2)$ percentile point from the normal distribution using a non-directional hypothesis test.

Roussos and Stout (1996b, p. 220) proposed the following \widehat{b}_{UNI} values for classifying DIF on a single item: (a) negligible or A-level DIF: Null hypothesis is rejected and the absolute value of $\widehat{b}_{UNI} < 0.059$, (b) moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\widehat{b}_{UNI}| < 0.088$, and (c) large or C-level DIF: Null hypothesis is rejected and $|\widehat{b}_{UNI}| \geq 0.088$.

Unfortunately, there are no comparable guidelines for classifying \widehat{b}_{UNI} when bundles are assessed.

To summarize, Roussos and Stout (1996a) proposed a DIF analysis paradigm that unifies substantive and statistical analyses by linking both to the Shealy-Stout (1993) multidimensional model of DIF. It consists of developing and testing DIF hypotheses using a combination of substantive and statistical analyses where the substantive analysis is used to develop the DIF hypotheses and the statistical analysis is used to test the hypotheses and estimate the amount of DIF. By combining substantive and statistical DIF analyses, researchers can begin to study the sources of DIF and to create a body of confirmed DIF hypotheses. These hypotheses, in turn, could lead to new policies, processes, and practices during the test construction, translation, and adaptation stages that might reduce the number of items that do not function equivalently across language groups. It should also lead to a better understanding of why DIF occurs.

Method

Student Sample and Achievement Tests

Data from eight different student samples in two content areas at two grade levels were analyzed in this study. At Grade 6, data from 3000 English and 2115 French Immersion students

who wrote the 1997 administration of a Mathematics and Social Studies Achievement Test were used. At Grade 9, data from 3000 English and 2115 French Immersion students who wrote the 1997 administration of a Mathematics and Social Studies Achievement Test were used. The samples in each content area at each grade level were randomly-selected from a database containing approximately 38000 English- and 3000 French-speaking students. The achievement tests were administered in the Canadian province of Alberta. Because Canada has two official languages, English and French, students have the option of obtaining their schooling in either language. In this study, the English-speaking examinees represent the dominant language group because the majority of students receive instruction in this language at English schools. English-speaking students are tested in English. Alternatively, the French Immersion students are in programs where French is the main language of instruction. Immersion students are tested in French.

In Grade 6, the mathematics test contained 50 multiple-choice items and each item had four options. Test items were classified into five curricular content areas: number relations, fractions, computation and operations, measurement and geometry, and data analysis. The social studies test contained 50 multiple-choice items, and each item had four options. Test items were classified into four curricular content areas: local government, ancient greek civilization, China, and geography and mapping.

In Grade 9, the mathematics test contained 43 multiple-choice items and six numeric response items that were intended to be equivalent across both languages. Each multiple-choice item had four options. Test items were classified into four curricular content areas: number, patterns and relations, shape and space, and statistics and probability. The social studies test contained 55 multiple-choice items, and each item had four options. Test items were classified into four curricular content areas: technology and change, economic systems, quality of life available from different economic systems, and the former USSR. For all four tests, items were based on concepts, topics, and facts from the province-wide Program of Studies (Alberta Education, 1989, 1996).

All items were developed in English by a committee of item writers and then translated into French using a four-step process. First, the items were translated from English to French by one translator during item development. The translator made reference to the Program of Studies and approved textbooks for grade level and subject specific terminology. Second, the translated test was validated by a committee comprising at least two French teachers along with a bilingual test developer. In this step, the comparability of the English and French version of the test was assessed using professional judgment. The validation committee also referred to the Program of Studies and to appropriate textbooks during the validation step. Once the committee had reviewed the test, the translator and test developer received comments and feedback on the accuracy and appropriateness of the translated test. Third, the test developer, acting on the recommendations of the committee, decided on the final changes. These changes were made by the translator and the translated test was finalized. Fourth, both the test developer and the test development supervisor reviewed and finalized the translated test. The translator in this process was a former teacher with 23 years experience in English-to-French translation.

Substantive Analysis

The substantive analysis began with a comprehensive review of DIF items from a 1996 administration of Grade 6 Mathematics and Social Studies Achievement Tests. All DIF items were identified with SIBTEST using a single-item analysis (i.e., studying one item at a time and using the remaining items as the matching subtest). B- and C-level DIF items were included in this review. An 11-member review committee was formed. The committee included three bilingual test translators, one bilingual test editor, two monolingual English-speaking psychometricians, two bilingual test developers, one monolingual English-speaking test developer, and two monolingual English-speaking assistant directors for test development. The committee members were shown the English and French test items. For each item, they were asked to identify any translation problems or differences, describe the source of this translation difference, and specify which group the item would favor. Discussion continued until consensus was reached on these three issues. The item review required six hours across two different

meetings. Four sources of translation DIF were identified in this review (the four sources, presented in Table 2, will be discussed in the Results section).

To validate the sources of translation DIF generated by the 11-member committee, two certified translators who had extensive experience translating educational tests, texts, and documents used the sources of translation DIF to identify the probable cause and direction of DIF for items from a 1997 administration of Grade 6 and 9 Mathematics and Social Studies Achievement Tests. All DIF items were identified with SIBTEST using a single-item analysis. B- and C-level DIF items were included in the review. One translator was a native English-speaker and the second was a native French-speaker. Both translators were completely bilingual in the English and French languages and both translators were certified by the Association of Translators and Interpreters of Alberta, which is an association affiliated with the Canadian Translators and Interpreters Council (CTIC) and the International Federation of Translators. To be a certified translator, applicants must pass the national CTIC exam once every three years.

The translators first independently identified any translation problem or difference, described the nature of this translation difference using the four sources outlined by the 11-member committee, and specified which group the item would favor. In other words, the translators categorized the items into eight bundles (4 sources by 2 groups) for each content area in both grade levels. Then, the translators met to discuss their decisions and to reach consensus on the items where they disagreed. The independent item review across both translators required 22 hours and the discussion until consensus required five hours. The Grade 6 and 9 items were reviewed in two different meetings¹.

Statistical Analysis

Once the translators' substantive reviews were completed, the item or bundle was tested. SIBTEST was run on the data from English and French examinees in each content area for both grade levels. All hypotheses were tested with a directional hypothesis test using the translators' expectations about which group would be favored. Items with no identifiable source of translation DIF were placed in a separate category and tested with a non-directional hypothesis test. All statistical tests were conducted with an alpha level of .01. The matching subtests for these

analyses were the non-DIF items. To purify the matching variable in each analysis, standard DIF analyses were conducted using the matching subtest items (i.e., items without DIF). All items produced negligible DIF results (i.e., all items were A-level) except for two in Grade 9 social studies. These two items were removed from the Grade 9 social studies matching subtest for the item and bundle analyses reported in the next section.

Results

Psychometric Characteristics of the Achievement Tests

A summary of the observed psychometric characteristics on the Mathematics and Social Studies Achievement Tests for the Grades 6 and 9 English and French examinees is presented in Table 1. The large samples were used in this study resulted in many statistically significant differences between groups that were not practically significance. Therefore, statistical outcomes are not presented but some general trends are noted. First, the psychometric characteristics of the items were comparable between the English- and French-speaking examinees. Second, the measures of internal consistency, difficulty (see, especially, the range of item difficulty), and discrimination were quite similar for both language groups in mathematics and social studies. The mean for the French examinees was higher than the mean for the English examinees on all tests except for Grade 6 social studies. Third, the standard deviations, skewness, and kurtosis were similar between the two groups for each test indicating that the test score distributions were comparable across language groups. Fourth, the number of words on the French forms was noticeably larger than the English forms, especially in social studies.

Insert Table 1 about here

Sources of Translation DIF

The 11-member test development and analysis committee identified four sources of translation DIF during their review of the 1996 achievement tests. All four sources were expected to affect the performance for one group of examinees. These sources were described as: (a) omissions or additions that affect meaning, (b) differences in the words, expressions, or sentence structure of

items that are inherent to the language and/or culture, (c) differences in the words, expressions, or sentence structure of items that are not inherent to the language and/or culture, and (d) differences in item format. The four sources of translation DIF are presented in Table 2.

Insert Table 2 about here

Similarities and differences between the sources of translation DIF identified by the 11-member committee and by Allalouf et al. (1999) are apparent. Source 1 in the current study, omissions or additions that affect meaning, is similar to “changes in content” identified by Allalouf et al. because both sources focus on changes in meaning resulting from a change to the item content. Source 2, differences in the words, expressions, or sentence structure of items that are inherent to the language and/or culture, is similar to “differences in culture relevance” from Allalouf et al. because both sources focus on inherent cultural differences apparent in both language groups that may affect student performance. Source 2 also differed from the Allalouf et al. categories in two ways. First, it included inherent language differences, in addition to cultural differences, that could affect student performance. Second, it overlapped with another cause described by Allalouf et al. as “changes in the difficulty of words or sentences” since both sources focus on the difficulty of words and sentences. The 11-member committee chose to differentiate between differences in words, expressions, and sentence structure inherent to language and/or culture (source 2) from differences not inherent to language and/or culture (source 3) whereas Allalouf et al. did not make this distinction. Source 3, differences in the words, expressions, or sentence structure of items that are not inherent to the language and/or culture, is similar to “changes in content” from Allalouf et al. because both sources identify changes in meaning that result in a potentially different item across the two language forms. Sources 1 and 3, while both tied to content and meaning, differentiate between omissions or additions compared to differences in words, expressions, and sentence structure not inherent to language and/or culture. Allalouf et al. did not make this distinction. Source 4, differences in punctuation, capitalization, item structure, typeface, and other formatting usages, is comparable to the source

described by Allalouf et al. as “changes in format” because both sources focus on item format differences.

These results demonstrate that common sources of translation DIF are found across language groups. Similar sources were found when the current study was compared to the Allalouf et al. (1999) study even though the studies were conducted independently with different language groups, content areas, and grade levels. Differences between the structure and organization of the translation DIF sources across the two studies were also noted.

Results from Certified Translators By Content Area

The substantive and statistical results for the 1997 Grade 6 and 9 Mathematics Achievement Test are presented in Tables 3 and 4, respectively. The Grade 6 Mathematics Achievement Test contained seven DIF items. The translators categorized five of these seven DIF items into three sources. The translators correctly predicted the group that would be favored for three of the four items or bundles, and these three predicted outcomes were statistically significant using a directional hypothesis test. They correctly anticipated the effect of source 1 items for the English-speaking examinees but not the French-speaking examinees. The translators correctly anticipated the effect of source 2 and 3 items on student performance (favoring French- and English-speaking students, respectively). The translators could not interpret two items from the Grade 6 test (labeled “No interpretation sources of translation DIF” in Table 4). When tested with a non-directional hypothesis test, the bundle was statistically significant and, overall, it favored the French examinees.

The Grade 9 Mathematics Achievement Test contained 11 DIF items, and all 11 were categorized into three sources. The translators correctly predicted the group who would be favored for all four items or bundles, and all predicted outcomes were statistically significant using a directional hypothesis test.

Insert Tables 3 and 4 about here

The results for the 1997 Grade 6 and 9 Social Studies Achievement Test are presented in Tables 5 and 6, respectively. The Grade 6 Social Studies Achievement Test contained 27 DIF items. The translators categorized these 27 DIF items into four sources of translation DIF. The translators correctly predicted the group who would be favored for four of the seven items or bundles, and all four outcomes were statistically significant using a directional hypothesis test. The translators did not correctly predict the effects of source 1 items for either language group. They correctly predicted the effects for source 2 items for French examinees but not English examinees. The translators predicted the effects of source 3 items for both language groups. They also predicted the effect of source 4 items for the English examinees. The translators could not interpret five items in Grade 6. When tested with a non-directional hypothesis test, the bundle was not statistically significant.

The Grade 9 Social Studies Achievement Test contained 17 DIF items. The translators categorized the DIF items into three sources. The translators correctly predicted the group who would be favored for four of the six items or bundles, and all four predicted outcomes were statistically significant using a directional hypothesis test. They anticipated the effect of source 1 items for both language groups. Conversely, they could not predict the effects of source 2 items for either language group. The translators anticipated the effect of source 3 items for both language groups. The translators could not interpret five items in Grade 9. When tested with a non-directional hypothesis, the bundle was not statistically significant.

Insert Tables 5 and 6 about here

Three outcomes are apparent when the mathematics and social studies results are compared. The Mathematics tests contained fewer DIF items indicating that translation differences were less pronounced in this content area (7 and 11 versus 27 and 17 for Grade 6 and 9 mathematics and social studies, respectively). The translators' predictions were more consistent with student outcomes in mathematics (3 of 4 and 4 of 4 versus 4 of 7 and 4 of 6 for Grade 6 and 9 mathematics and social studies, respectively). The Mathematics tests produced fewer

uninterpretable items for the translators (2 and 0 versus 5 and 5 for Grade 6 and 9 mathematics and social studies, respectively).

Confirmed DIF Hypotheses

The main purpose of this study is to identify sources of differential item functioning on translated achievement tests using substantive and statistical DIF analyses, as described by Roussos and Stout (1996a). By taking a confirmatory approach, researchers can begin to study the sources of DIF and to create a body of confirmed DIF hypotheses. The outcomes in mathematics were, for the most part, interpretable. The translators categorized the DIF items into three sources. They correctly predicted the group that would be favored for seven of the eight items or bundles across the two grade levels, and these seven predicted outcomes were statistically significant using a directional hypothesis test. Only one bundle, containing two items, was incorrectly predicted. The majority of items across the two grade levels were associated with source 3, differences in the words, expressions, or sentence structure of items that are not inherent to the language and/or culture, and the translators correctly anticipated the effects of these translations differences for both English and French examinees. The translators could not interpret two items from the Grade 6 Mathematics test and, when tested with a non-directional hypothesis test, the bundle was statistically significant. This outcome suggests these items produced a systematic effect that favored the French examinees but the source of this effect could not be identified by the translators. This outcome also indicates that the four sources accounted for most of the translation differences identified by the translators on the Mathematics tests.

The outcomes in social studies were more complex and less interpretable compared to mathematics. The translators categorized the social studies DIF items into four sources. They correctly predicted the group who would be favored for eight of the 13 items or bundles across the two grade levels, and these eight predicted outcomes were statistically significant using a directional hypothesis test. Source 1 bundles favored the predicted language group in Grade 9 but not in Grade 6. The outcomes for the source 2 bundles were less consistent. Only one of the four bundles were correctly predicted by the translators—those items favoring the French

examinees in Grade 6. Take together, the items for these two sources of translation DIF in social studies were not consistently associated with student performance. This outcome could be attributed to either the improper identification of items by the translators (i.e., the items did not have translation differences associated with source 1 or 2) or the inaccurate prediction of student performance (i.e., the translators expected that the difference would favor English examinees when, in fact, it favored the French examinees). Source 3 bundles favored the predicted language group in both grades. Moreover, the majority of the social studies items were associated with source 3 DIF across the two grade levels, as with mathematics, and the translators correctly anticipated the effects of these translations differences for both language groups. The four items associated with the source 4 bundle in Grade 6 also favored the predicted language group. Translation DIF associated with source 4, format differences, while rare, was reliably identified and predicted by the translations. The translators could not interpret 10 DIF items from the two social studies tests. When tested with a non-directional hypothesis test, the bundles were not statistically significant suggesting that the five items in each bundle do not systematically favor one language group. This finding suggests that factors other than those listed in Table 2 account for differential performance. As such, the four sources are less effective at explaining translation differences in social studies compared to mathematics.

Conclusions and Discussion

The psychometric literature contains many statistical methods for identifying DIF (see reviews by Clauser & Mazor, 1998; Millsap & Everson, 1993). Despite the presence of these methods, many researchers and practitioners agree that identifying the sources of these differences is difficult, even when DIF statistics are used. In an attempt to overcome this problem in the area of test translation and adaptation, substantive and statistical analyses were used to identify the sources of differential item functioning on translated achievement tests. A substantive analysis of existing DIF items was conducted by an 11-member committee of test developers, analysts, editors, and translators to produce a list with sources of translation DIF (see Table 2). Then, two certified translators used this list to categorize DIF items from Grade 6 and 9 Mathematics and Social Studies Achievement Tests into four different sources. Finally, the items or bundles, each

associated with a specific source of translation DIF and each anticipated to favor a specific group of examinees, were tested statistically using SIBTEST. This approach is consistent with the Roussos-Stout (1996a) DIF analysis paradigm that attempts to unify substantive and statistical analyses by linking both to the Shealy-Stout (1993) multidimensional model of DIF. By combining substantive and statistical DIF analyses, researchers can study the sources of translation DIF and create a body of confirmed DIF hypotheses that may influence testing policies and practices and provide researchers and practitioners with a better understanding of why DIF occurs.

The sources of translation DIF identified by an 11-member committee of test developers, analysts, editors, and translators appear to be robust since they were used by two translators to account for differences on English tests translated into French in two diverse content areas at two grade levels. However, the results were not uniform across content area. The translators' predictions were more consistent with student outcomes in mathematics compared to social studies. In mathematics, seven of eight bundles created by the certified translators using pre-specified translation DIF sources produced significant \hat{b} values. In social studies, eight of 13 bundles created by the translators produced significant \hat{b} values suggesting that the cause of translation DIF for these items was more difficult to identify owing, perhaps, to the larger number of DIF items and the difference in complexity of the two items types (e.g., number of words, connotations, sentence structure, typically of words and phrases, etc. differ between the social studies and mathematics items).

When the results are examined across content area and grade level, source 3 bundles, associated with differences in the words, expressions, or sentence structure of items that are not inherent to the language and/or culture, were consistently identified in both content areas at both grade levels. These bundles also contained the largest number of DIF items. It appears that this source of translation DIF is common and can be reliably identified by skilled translators. Translators in the test development process should be aware of the negative effects that differences in the words, expressions, or sentence structure of items can produce on the comparability of items across two language forms and remove these differences. The source 4

bundle, associated with differences in punctuation, capitalization, item structure, typeface, and other formatting usages, was only found once in this study but, when identified, produced a statistically significant outcome in favor of the predicted language group (the English examinees on the Grade 6 Social Studies Achievement Test). Clearly, item format must be consistent across language forms, and this consistency should be monitored during test development.

Limitations

Social studies contained 10 items that were not associated with the four sources of translation DIF while mathematics had only two such items. Roussos and Stout (1996a), in the MMD framework, claim that the secondary dimensions (i.e., the dimensions that cause DIF) are considered auxiliary if they are intentionally assessed or nuisance if they are unintentionally assessed as part of the construct on the test. DIF that is caused by auxiliary dimensions is benign whereas DIF that is caused by nuisance dimensions is adverse. In the current study, we were trying to identify and interpret adverse DIF associated with translation differences. These differences produce non-equivalent items across language forms. Moreover, this type of DIF, if identified early in the test development process, can be minimized or even eliminated. Most translation differences identified in mathematics were attributed to three sources of adverse DIF. Social studies, on the other hand, contained 10 DIF items that were not classified using the four sources in this study. In addition, five of the 13 bundles were not correctly predicted. This outcome could be attributed to either the improper identification of translation DIF by the translators or the inaccurate prediction of student performance. Research must now be undertaken to resolve this issue. Could these outcomes be attributed to translation differences that were overlooked by the translators? Or, are these outcomes a result of genuine performance differences between English and French examinees? Clearly, translation differences were more pronounced in social studies. This outcome might be related to the disproportionate number of words on each form (i.e., 3354 to 4157 and 4490 to 5247 on the English and French forms of the social studies test at Grade 6 and 9, respectively) which could introduce secondary dimensions, such as reading comprehension, vocabulary knowledge, cognitive capabilities (e.g., information-processing speed related to recognizing and recalling the meaning of words), or testwiseness that

affect student performance. Performance differences might be interpreted as benign if they are considered a logical component of the social studies construct. They could also be interpreted as adverse if any of these factors are considered nuisance dimensions. To interpret and judge the nature of these differences, researchers must focus on the psychological dimensions that produce translation DIF.

Future Directions

Many researchers and practitioners in the psychometric community agree that the psychology of test performance must be understood to develop, score, and validly interpret results from educational tests (e.g., Frederiksen, Mislevy, & Bejar, 1993; Gierl, 1997; Hattie, Jaeger, & Bond, 1999; Mislevy, 1996; Nichols, 1994; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1989; Wiley, 1991). Consensus on this view seems inevitable because most educational tests are based on cognitive problem-solving tasks. Yet, little is known about the cognitive processes actually used by examinees as they respond to achievement test items in different languages. The next step is to develop more refined and detailed DIF hypotheses where researchers identify content differences and predict group differences (as was done in the current study) but also study the content-by-group interactions (Cronbach, 1989, p. 155). The last step requires a better understanding of the psychology underlying test performance where examinees' cognitive processes are studied in context as they respond to test items using methods such as protocol analysis (Ericsson & Simon, 1993) rather than relying on content specialists, psychometricians, or translators to make inferences about examinees' cognitive skills.

An example helps illustrate this point. One item from the 1997 Grade 9 Mathematics test is presented in the Appendix. Examinees were shown a diagram outlining the floor plan of a bedroom and asked to compute the longest straight line that could be painted on the ceiling. The English stem reads, 'If Ella wanted to paint a stripe on the ceiling of her room, what is the longest straight line she could paint?'. The French stem reads, 'Si Ella voulait peindre une bande sur le plafond de sa chambre, quelle est la bande en ligne droite la plus longue qu'elle pourrait peindre?' The French form has a parallel structure across the two phrases emphasizing the word 'bande' ('band'). The English form does not have this structure using the word 'stripe' in the first

part of the phrase and 'straight line' in the second part. One translator argued this item would favor the French examinees because the parallel structure makes the question clearer in French by relating 'voulait peindre une bande' ('wanted to paint a band') to 'quelle est la bande en ligne droite la plus longue' ('what is the longest straight band'). The other translator believed this item would favor the English examinees despite the parallel structure because the phrase 'quelle est la bande en ligne droite la plus longue' ('what is the longest straight band') is an unusual expression that would be unfamiliar to many French examinees compared to the more familiar English phrase 'what is the longest straight line'. These competing interpretations of DIF both sound plausible. To resolve this disagreement, a protocol analysis focusing on the strategies and cognitive processes actually used by examinees could be invaluable because it may provide evidence about how students interpret and solve this problem. (The math item in this example favored the English examinees.)

By combining the results from judgmental reviews and student protocols in a confirmatory framework, we move one step closer to understanding the psychology of DIF using principles in cognitive psychology and practices in educational measurement to identify and interpret the dimensions that affect student performance. It is an approach advocated by Messick (1989) when he wrote:

Almost any kind of information about a test can contribute to an understanding of its construct validity, but the contribution becomes stronger if the degree of fit of the information with the theoretical rationale underlying score interpretation is explicitly evaluated....Possibly most illuminating of all are direct probes and modeling of the processes underlying test responses, an approach becoming both more accessible and more powerful with continuing developments in cognitive psychology (p. 17).

Researchers must heed this suggestion by studying the content and cognitive dimensions on translated tests. The outcomes from this type of research will suggest new test development practices, provide explanations about the nature of group differences, and help us understand the sources of DIF on translated and adapted tests.

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.
- Alberta Education. (1989). Program of Studies: Social Studies. Edmonton, AB: Curriculum Standards Branch, Alberta Education.
- Alberta Education. (1996). Program of Studies: Mathematics. Edmonton, AB: Curriculum Standards Branch, Alberta Education.
- 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. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.
- Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins Press.
- Bond, L. (1993). Comments on the O'Neill and McPeck paper. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 277-279). Hillsdale, NJ: Lawrence Erlbaum.
- Camilli, G., & Shepard, L. (1994). Methods for identifying biased test items. Newbury Park: Sage.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice, *17*, 31-44.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence: Measurement theory and public policy (Proceedings of a symposium in honor of Lloyd G. Humphreys, pp. 147-171). Urbana, IL: University of Illinois Press.
- Douglas, J., Roussos, L., and Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. Journal of Educational Measurement, *33*, 465-484.
- Englehard, 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. (1999, April). Translation DIF on TIMSS. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data (Rev. ed.). Cambridge, MA: MIT Press.

Frederiksen, N., Mislevy, R. J., Bejar, I. I. (1993). Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. Journal of Educational Research, *91*, 26-32.

Gierl, M., & McEwen, N. (1998, May). Differential item functioning on the Alberta Education Social Studies 30 Diploma Exams. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Ottawa, Ontario, Canada.

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. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, *10*, 229-244.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. Social Indicators Research, *45*, 153-171.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. Review of Research in Education, *24*, 393-446.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. European Journal of Psychological Assessment, *15*, 277-283.

Jiang, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. Journal of Educational & Behavioral Statistics, *23*, 291-322.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: American Council on Educational, Macmillian.

- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 297-334.
- Mislevy, R. J. (1996). Test theory reconceived. Journal of Educational Measurement, 33, 379-416.
- Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. Review of Educational Research, 64, 575-603.
- Nichols, P. D., Chipman, S. F., Brennan, R. L. (1995). Cognitively diagnostic assessment. Hillsdale, NJ: Erlbaum.
- O'Neill, K. A., & McPeck, 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.
- Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. Educational and Psychological Measurement, 40, 397-404.
- Ramsey, P. A. (1993). Sensitivity reviews: The ETS experience as a case study. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum.
- Reckase, M. D., & Kunce, C. (1999, May). Translation accuracy of a technical credentialing examination. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.
- Rengel, E. (1986, August). Agreement between statistical and judgmental item bias methods. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.
- Roussos, L. A., & Stout, W. F. (1996b). 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.

Sandoval, J. ., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58, 159-194.

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

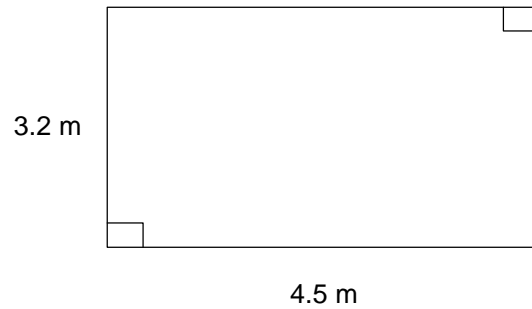
Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 263-331). New York: American Council on Educational, Macmillian.

Stout, W., & Roussos, L. (1995). SIBTEST Manual. Urbana, IL: University of Illinois, Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), Improving inquiry in social science (pp. 75-107). Hillsdale, NJ: Erlbaum.

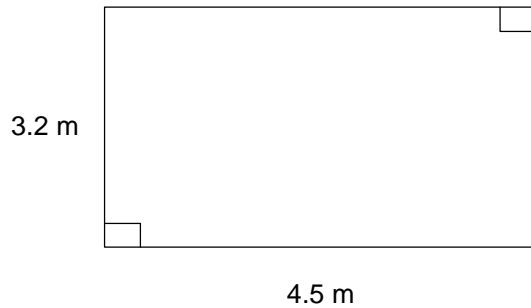
Appendix

The diagram below shows the floor plan of Ella's bedroom.



2. If Ella wanted to paint a stripe on the ceiling of her room, what is the longest straight line she could paint?
- A. 4.5 m
 - B. 5.5 m
 - C. 7.7 m
 - D. 15.4 m

Le diagramme ci-dessous montre le plan de la chambre à coucher de Ella.



2. Si Ella voulait peindre une bande sur le plafond de sa chambre, quelle est la bande en ligne droite la plus longue qu'elle pourrait peindre?
- A. 4.5 m
 - B. 5.5 m
 - C. 7.7 m
 - D. 15.4 m

Author Notes

Please address all correspondence to Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, 6-110 Education Centre North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5. Email: mark.gierl@ualberta.ca.

This research was supported with funds awarded to the first author from the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Social Science Research Fund (SSR) at University of Alberta.

We would like to thank Jeffrey Bisanz for his helpful comments on an earlier version of this manuscript.

Footnote

¹Our approach is a slight departure from the substantive method suggested by Roussos and Stout (1996a) because our translators only reviewed items with B- or C-level DIF instead of reviewing all the items on the test. We made this decision because four tests were evaluated in this study and a large number of DIF items were identified. By focusing on the DIF items, the substantive review, which was both expensive and time consuming, became a more manageable task.

Table 1

Psychometric Characteristics for the Four Achievement Tests

Characteristic	<u>Grade 6</u>				<u>Grade 9</u>			
	<u>Mathematics</u>		<u>Social Studies</u>		<u>Mathematics</u>		<u>Social Studies</u>	
	English	French	English	French	English	French	English	French
No. of Examinees	3000	2115	3000	2115	3000	2115	3000	2115
No. of Items	50	50	50	50	49	49	55	55
No. of Words	2713	3066	3354	4157	3118	3179	4490	5247
Mean	35.30	36.56	33.68	32.17	29.04	33.52	38.31	39.04
SD	8.46	7.45	8.34	7.77	10.24	8.72	8.98	7.98
Skewness	-0.49	-0.67	-0.45	-0.32	-0.16	-0.42	-0.56	-0.63
Kurtosis	-0.47	-0.10	-0.39	-0.47	-0.90	-0.51	-0.36	-0.08
Internal Consistency ^a	0.89	0.86	0.87	0.84	0.92	0.89	0.88	0.86
Mean Item Difficulty	0.71	0.73	0.67	0.64	0.59	0.68	0.70	0.71
SD Item Difficulty	0.15	0.16	0.12	0.12	0.15	0.15	0.14	0.15
Range Item Difficulty	0.26 – 0.90	0.17 – 0.93	0.40 – 0.87	0.39 – 0.88	0.25 – 0.91	0.39 – 0.98	0.38 – 0.93	0.33 – 0.92
Mean Item Discrimination ^b	0.48	0.44	0.43	0.37	0.54	0.50	0.46	0.41
SD Item Discrimination	0.12	0.14	0.11	0.11	0.14	0.12	0.13	0.14
Range Item Discrimination	0.02 - 0.68	0.11 - 0.67	0.17 – 0.65	0.16 - 0.58	0.29 – 0.79	0.26 – 0.76	0.17 – 0.68	0.03 – 0.66

^aCronbach's alpha^bBiserial correlation

Table 2

Sources of Translation Differential Item Functioning as Identified by a 11-Member Committee of Test Translators, Editors, Analysts, and Developers

SOURCE #1—Omissions or Additions that Affect Meaning: Omissions or additions of words, phrases, or expressions that affect meaning and are likely to affect the performance for one group of examinees. For example, on an item with a contour relief map, the English form contained the phrase ‘cross section cut along a line’ while the French form contains the phrase ‘une coupe transversale qui montre le relief’. The idea of ‘relief’ is excluded from the English form. Another example: The English form contained the expression ‘this number written in standard form is’ while the French form had the phrase ‘ce nombre est’ (‘the number is’). The idea of ‘standard form’ is excluded from the French translation.

SOURCE #2—Differences in Words or Expressions Inherent to Language or Culture: Differences in the words, expressions, or sentence structure of items that are inherent to the language and/or culture and are likely to affect the performance for one group of examinees. One example to illustrate a language difference includes the English sentence, ‘Most rollerbladers favour a helmet bylaw.’ versus the French translation, ‘La plupart des personnes qui ne font pas de patin à roulettes sont pour un règlement municipal en faveur du port du casque protecteur.’ The expressions for rollerbladers (patin à roulettes) and helmet bylaw (un règlement municipal en faveur du port du casque protecteur) differ dramatically between English and French forms because, as noted by the translators, rollerblader and helmet bylaw have no expression that is directly parallel in French. One example to illustrate a cultural difference includes an English item with a 12-hour clock using AM and PM while the French translation uses a 24-hour clock. This convention, the 12- vs. the 24-hour clock, represents an English-French cultural difference.

SOURCE #3—Differences in Words or Expressions Not Inherent to Language or Culture: Differences in the words, expressions, or sentence structure of items that are not inherent to the language and/or culture and that are likely to affect the performance for one group of examinees. For example, the phrase in English ‘basic needs met’ versus the phrase in French ‘les services offerts’ focuses on ‘needs’ in English and ‘services’ in French. A second example: The English phrase ‘traditional way of life’ versus the French phrase ‘les traditions’ present two distinct concepts surrounding ‘a way of life’ and ‘traditions’ in the English and French forms, respectively. A third example: The English phrase ‘animal power’ versus the French phrase ‘à l’aide des animaux’ present distinct concepts related to ‘the power of animals’ and ‘the help of or aid by animals’ in the English and French forms, respectively. In all three examples, alternative

phrases were identified by the two translators in our study that would produce items that were closer in meaning across the two languages. Hence, these differences are not inherent to the languages unlike the examples in source #2.

SOURCE #4—Format Differences: Differences in punctuation, capitalization, item structure, typeface, and other formatting usages that are likely to affect the performance for one group of examinees. For example, a word that appeared only in the stem for the English form was presented in all four options for the French form thus representing a difference in item structure. As a result, the French item was more cumbersome. Similarly, an item contained a title in capital letters in one form but not the other representing a difference in typeface. If these differences provide a clue to the correct answer for one group of examinees, then the item may not be comparable across language groups.

Table 3

Results for the Grade 6 1997 Mathematics Achievement Test Administered in English and French

Source	Item No.	Favors (Predicted)	Beta-Uni
1—Omissions or Additions that Affect Meaning	6	English	0.084*
	41, 49	French	0.022
2— Differences in Words or Expressions Inherent to Language or Culture	47	French	-0.118*
3—Differences in Words or Expressions Not Inherent to Language or Culture	44	English	0.324*
4—Format Differences	---	---	---
No Identifiable Sources of Translation DIF	27, 39	---	-0.108*

* $p < .01$.Note. A negative SIB-Uni indicates the item or bundle favors the French examinees.

Table 4

Results for the Grade 9 1997 Mathematics Achievement Test Administered in English and French

Source	Item No.	Favors (Predicted)	Beta-Uni
1—Omissions or Additions that Affect Meaning	26, 40	English	0.184*
2— Differences in Words or Expressions Inherent to Language or Culture	27	English	0.098*
3—Differences in Words or Expressions Not Inherent to Language or Culture	2, 25, 27, 36, 40	English	0.470*
	15, 37, NR8	French	-0.252*
4—Format Differences	---	---	---
No Identifiable Sources of Translation DIF	---	---	---

* $p < .01$.Note. A negative SIB-Uni indicates the item or bundle favors the French examinees.

Table 5

Results for the Grade 6 1997 Social Studies Achievement Test Administered in English and French

Source	Item No.	Favor (Predicted)	Beta-Uni
1—Omissions or Additions that Affect Meaning	19, 46	English	-0.007
	45	French	0.078
2— Differences in Words or Expressions Inherent to Language or Culture	22, 36	English	-0.028
	5, 7, 13	French	-0.404*
3—Differences in Words or Expressions Not Inherent to Language or Culture	2, 8, 9, 11, 19, 24 25, 27, 33, 40, 47	English	0.952*
	3, 5, 16, 18, 35	French	-0.404*
4—Format Differences	17, 33, 44	English	0.310*
No Identifiable Sources of Translation DIF	6, 29, 30, 34, 48	--	-0.066

* $p < .01$.Note. A negative SIB-Uni indicates the item or bundle favors the French examinees.

Table 6

Results for the Grade 9 1997 Social Studies Achievement Test Administered in English and French

Source	Item No.	Favor (Predicted)	Beta-Uni
1—Omissions or Additions that Affect Meaning	9, 34, 52	English	0.170*
	7	French	-0.060*
2— Differences in Words or Expressions Inherent to Language or Culture	9	English	-0.169
	10, 12, 16, 24	French	0.055
3—Differences in Words or Expressions Not Inherent to Language or Culture	1, 6	English	0.241*
	2, 3, 7, 10, 12, 24	French	-0.299*
4—Format Differences	---	---	---
No Identifiable Sources of Translation DIF	21, 27, 39, 45, 48	---	-0.071

* $p < .01$.Note. A negative SIB-Uni indicates the item or bundle favors the French examinees.