The Alberta Journal of Educational Research *Vol. XLV, No. 4, Winter 1999, 353-376*

Mark J. Gierl

W. Todd Rogers

and

Don A. Klinger
University of Alberta

Using Statistical and Judgmental Reviews to Identify and Interpret Translation Differential Item Functioning

The purpose of this study was to evaluate the equivalence of two translated tests using statistical and judgmental methods. Performance differences for a large random sample of English- and French-speaking examinees were compared on a grade 6 mathematics and social studies provincial achievement test. Items displaying differential item functioning (DIF) were flagged using three popular statistical methods—Mantel-Haenszel, Simultaneous Item Bias Test, and logistic regression—and the substantive meaning of these items was studied by comparing the back-translated form with the original English version. The items flagged by the three statistical procedures were relatively consistent, but not identical across the two tests. The correlation between the DIF effect size measures were also strong, but far from perfect, suggesting that two procedures should be used to screen items for translation DIF. To identify the DIF items with translation differences, the French items were back-translated into English and compared with the original English items by three reviewers. Two of seven and six of 26 DIF items in mathematics and social studies respectively were judged to be nonequivalent across language forms due to differences introduced in the translation process. There were no apparent translation differences for the remaining items, revealing the necessity for further research on the sources of translation differential item functioning. Results from this study provide researchers and practitioners with a better understanding of how three popular DIF statistical methods compare and contrast. The results also demonstrate how statistical methods inform substantive reviews intended to identify items with translation differences.

Le but de cette étude était d'évaluer l'équivalence de deux examens traduits avec des méthodes basées sur les statistiques et d'autres reposant sur le jugement. On a comparé les différences dans la performance d'un grand échantillon aléatoire de sujets anglophones et francophones qui avaient complété des examen provinciaux de sixième année en mathématiques et en études sociales. Les items démontrant une divergence par rapport aux autres (differential item functioning - DIF) ont été marqués d'un indicateur dans le contexte de trois méthodes statistiques bien connues - Mantel-Haenszel, Simultaneous Item Bias Test et la régression

---

Mark Gierl is an assistant professor of educational psychology. His research interests include educational measurement and evaluation, with an emphasis on assessment and cognition, differential item functioning, and item response theory.
Todd Rogers is a professor in the Department of Educational Psychology and Director of the Centre for Research in Applied Measurement and Evaluation. He specializes in measurement and evaluation.
Don Klinger is a doctoral student in educational psychology. His research interests include educational measurement and school effectiveness, with an emphasis on testing issues and setting standards.

logistique. La signification de fond de ces items a été étudiée en comparant la version traduite de l'examen avec l'original en anglais. Les items marqués par les trois procédures statistiques étaient relativement constants mais pas identiques d'une version à l'autre. Alors que la corrélation entre les mesures de l'effet DIF étaient aussi forte, elle était loin d'être parfaite, ce qui suggère que l'on devrait avoir recours à deux procédures dans le dépistage du DIF en traduction. Pour identifier les items DIF présentant des différences en traduction, trois réviseurs ont comparé les items français retraduits en anglais avec les originaux en anglais. Ceux-ci ont jugé que deux sur sept items en mathématiques et six sur vingt-six items en études sociales n'étaient pas équivalents d'une langue à l'autre à cause des différences introduites par le processus de traduction. Les autres items ne présentaient pas de différences apparentes de traduction, ce qui révèle le besoin de poursuivre la recherche sur les sources du DIF en traduction. Les résultats de cette étude aideront les chercheurs et les praticiens à mieux comprendre les similarités et les différences entres trois méthodes statistiques DIF souvent employées. De plus, ils démontrent comment les méthodes statistiques contribuent aux études de signification dont le but est l'identification des items présentant des différences de traduction.

Item bias is a serious concern for test developers and test users. Bias results in systematic errors that distort the inferences made from a test for members of a particular group such as female, Native, or French-speaking examinees. Bias occurs when items are worded in such a way that examinees from a specific group who are knowledgeable about the construct of interest are prevented from demonstrating their knowledge. In most cases, test items are biased because they contain sources of difficulty that are irrelevant or extraneous to the construct being tested, and this difficulty factor adversely affects test performance (Camilli & Shepard, 1994).

Bias is also a concern when a test is translated or adapted from one language or culture to another (Allalouf & Sireci, 1998; Budgell, Raju, & Quartetti, 1995; Hambleton, 1994). For example, the meaning of an item can change during test translation. Hambleton provides one illustrative example. In a Swedish-English comparison, English-speaking examinees were presented with this item:

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

In the Swedish translation the phrase *webbed feet* became *swimming feet*, thereby providing an obvious clue to the Swedish-speaking examinees about the correct option for this item. This type of testing problem has an important and obvious consequence: A difference in student performance resulting from such an item may be attributed to a group difference in achievement that is unfounded because the item is not equivalent across the two language groups.

Differential Item Functioning, Item Bias, and Item Impact

Differential item functioning (DIF) is present when examinees from different groups have a different probability or likelihood of answering an item correctly, after controlling for overall ability (Shepard, Camilli, & Averill, 1981). Once identified, DIF may be attributed to item bias or to item impact. *Item bias* is defined as invalidity or systematic error in how a test item measures a construct for the members of a particular group (Camilli & Shepard, 1994). It is sys-

tematic because it constantly distorts performance for members of the group. When a test item unfairly favors one group of examinees over another, the item is biased. Alternatively, group disparity in item performance that reflects actual knowledge and experience differences on the construct of interest is called *item impact*. Impact is also constant for the members of a particular group, but these effects reflect performance differences that the test is intended to measure (Clauser & Mazor, 1998).

The link between DIF, bias, and impact is largely methodological: Statistical analyses are used to identify items with DIF, and judgmental analyses are used to determine if DIF is attributable to bias or to impact for members of a specific group. Reviews intended to evaluate the fairness of an item cannot proceed as either a statistical or a judgmental analysis; both procedures are needed (Linn, 1993; van de Vijver, 1994).

Purpose of Study

Statistical and substantive issues in test translation are addressed in this study. First, the results from three statistical methods designed to identify DIF are compared: Mantel-Haenszel, Simultaneous Item Bias Test, and logistic regression. Currently all three methods are popular, but few studies have compared the outcomes from these procedures using a large sample of student response data from a major testing program. In this study, English- and French-speaking examinees who wrote a grade 6 mathematics and social studies achievement test are compared. A new DIF effect size measure used with the logistic regression procedure, as proposed by Zumbo and Thomas (1996), is also evaluated by comparing the results from this measure against existing effect size measures used with the Mantel-Haenszel procedure and the Simultaneous Item Bias Test.

Second, the utility of back-translation as a judgmental method for interpreting DIF is evaluated. Recall that the distinction between DIF, item bias, and item impact is important because DIF is a statistical concept and bias and impact are substantive concepts. Judgmental reviews intended to evaluate the equivalence of a test should rely on both statistical and judgmental analyses. In this study, back-translation is used to identify items that differ in meaning between the English and French forms of the tests. There is no attempt to account for item impact. Judgmental reviews that yield interpretable results are *essential* for identifying items with translation differences and for controlling this problem in future forms of the test. The efficacy of back-translation for achieving this outcome is evaluated by comparing the results from two back-translations of the same test against the original test and against the statistical outcomes.

Overview of Statistical and Judgmental Procedures

Statistical and judgmental procedures are used to establish the equivalence between source (i.e., original) and target (i.e., translated) language tests. In the first section, the three popular statistical procedures are briefly reviewed and compared. In the second section, the back-translation procedure is reviewed and a modified design is presented.

Mantel-Haenszel

Mantel-Haenszel (MH) is a nonparametric approach for identifying DIF (Holland & Thayer, 1988; Mantel & Haenszel, 1959). MH yields a chi-square test with one degree of freedom to test the null hypothesis that there is no relation between group membership and test performance on one item after controlling for ability. The MH procedure is also used to estimate the constant odds ratio that yields a measure of effect size for evaluating the amount of DIF. MH is computed by matching examinees in each group on total test score and then forming a 2-by-2-by-*K* contingency table for each item, where the score is level on the matching variable of total test score.

Research at the Educational Testing Service has resulted in proposed values for classifying the DIF effect size measure at the item level. DIF is considered negligible when the effect size measure Δ_{MH} is not significantly different from 0 and the magnitude of the $|\Delta_{MH}| < 1$. DIF is considered moderate when Δ_{MH} is significantly different from 0 and $1 \leq |\Delta_{MH}| < 1.5$. DIF is considered large when Δ_{MH} is significantly greater than 0 and $|\Delta_{MH}| \geq 1.5$. (Zieky, 1993; Zwick & Ercikan, 1989). These ratings are referred to as A-, B-, and C-level DIF to denote negligible, moderate, and large amounts of DIF.

Simultaneous Item Bias Test

The Simultaneous Item Bias Test (SIBTEST) is an alternative statistical method for detecting DIF proposed by Shealy and Stout (1993). SIBTEST is intended to model multidimensional data, although it can be used for unidimensional data as well. The statistical hypothesis tested by SIBTEST is:

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

where $B(T)$ is the difference in probability of a correct response on the studied item for examinees in the Reference and Focal groups matched on true score; $P_R(T)$ is the probability of a correct response on the studied item for examinees in the Reference group with true score T; and $P_F(T)$ is the probability of a correct response on the studied item for examinees in the Focal group with true score T. In other words $B(T)$, the parameter representing the amount of unidimensional DIF when a single test item is evaluated, is 0 when there is no DIF and nonzero when DIF is present. With the SIBTEST approach, items on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest contains the biased item, and the matching subtest contains the rest of the items. For each matching subtest score k, the corresponding subtest true score for the Reference and Focal groups, is estimated using linear regression. The estimated true scores are then adjusted using a regression correction technique to ensure that the estimated true score is comparable for the examinees in the Reference and Focal groups on the matching subtest. In the final step, $B(T)$ is estimated using $\hat{B}$, which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels.

Like the MH procedure, SIBTEST yields an overall statistical test as well as a measure of the effect size for each item ($\hat{B}$ is an estimate of the amount of DIF). Roussos and Stout (1996) suggested a range of values for interpreting $\hat{B}$. When

the null hypothesis is rejected and $|\hat{B}| < 0.059$, DIF is considered negligible; when the null hypothesis is rejected and $0.059 \leq |\hat{B}| < 0.088$, DIF is considered moderate; and when the null hypothesis is rejected and $|\hat{B}| \geq 0.088$, DIF is considered large. These guidelines are used to classify items in category A, B, or C (i.e., negligible, moderate, and large amounts of DIF).

SIBTEST differs from MH in a number of ways. First, SIBTEST uses a regression estimate of the true score instead of an observed score as the matching variable. As a result, examinees are matched on a latent rather than an observed score. Second, SIBTEST can be used to evaluate DIF in two or more items simultaneously in the analysis. This feature allows the developer to assess DIF more effectively in testlets or item bundles on an test (Douglas, Roussos, & Stout, 1996). Third, SIBTEST can be used to assess DIF iteratively by initially using all items in the matching variable and then systematically removing DIF items from the matching test until a subtest of items without DIF is identified. SIBTEST performs similarly to MH in identifying uniform DIF even with small samples of examinees (Narayanan & Swaminathan, 1994; Roussos & Stout, 1996).

Logistic Regression

A third approach commonly used to identify DIF is logistic regression (LR) (Swaminathan & Rogers, 1990). LR can detect both *uniform* and *nonuniform* DIF, which provides a distinct advantage over MH and SIBTEST as these two procedures were designed to detect only uniform DIF. Uniform DIF exists when there is no interaction between ability level and group membership. That is, the probability of answering an item correctly is greater for one group uniformly over all ability levels. In item response theory terminology, uniform DIF is indicated by parallel item characteristic curves. Nonuniform DIF occurs when there is an interaction between ability level and group membership. In this case the difference in the probabilities of a correct response for the two groups is not the same at all levels of ability. Simulation studies have been conducted to demonstrate that LR is in fact more powerful than MH and SIBTEST at detecting nonuniform DIF (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993).

The presence of DIF in the LR approach is determined by testing the improvement in model fit that occurs when a term for group membership and a term for the interaction between test score and group membership are successively added to the regression model. A chi-square test is then used to evaluate the presence of uniform and nonuniform DIF on the item of interest by testing each term included in the model. The general model for logistic regression takes the form

$$P(u=1) = \frac{e^z}{1+e^z},$$

where u is the score on the studied item. Performance on the studied item is first conditioned on total test score. In this step, $z = \beta_0 + \beta_1 X$, where X is the test score (Model 1). This serves as the baseline model. The presence of uniform DIF is then tested by examining the improvement in chi-square model fit associated with adding a term for group membership (G) against the baseline model, that

is, Model 1 subtracted from Model 2 ($z = \beta_0 + \beta_1 X + \beta_2 G$). The presence of nonuniform DIF is tested by examining the improvement in chi-square model fit associated with adding a term for group membership (G) *and* a term for the interaction between test score and group membership (XG) against Model 2, in other words, Model 2 subtracted from Model 3 ($z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$). Nonuniform DIF can be tested with LR regardless of the outcome from the uniform DIF test because each model contains different terms.

Zumbo and Thomas (1996) developed an index to quantify the magnitude of DIF for the LR procedure based on partitioning a weighted least-squares estimate of R^2 that yields an effect size measure (also see Pope, 1997; Thomas & Zumbo, 1996; Zumbo, 1999). This index is obtained first by computing the R^2 measure of fit for each term in the LR model (i.e., test score, group membership, test score-by-group membership interaction) and then by partitioning the R^2 for each of the terms. A DIF effect size for the group membership term is produced by subtracting the R^2 for the total test score term (Model 1) from the R^2 for the group membership term (Model 2). The result is an effect size measure associated with group membership that quantifies the magnitude of uniform DIF (herein called $R^2\Delta$–U). A second DIF effect size is produced for the total score-by-group membership term by subtracting the R^2 for the group membership term (Model 2) from the R^2 for the total score-by-group membership interaction term (Model 3). The result is an effect size measure associated with the total score-by-group membership interaction that quantifies the magnitude of nonuniform DIF (herein called $R^2\Delta - N$). As with the MH and SIBTEST effect size measures, $R^2\Delta$ can be used with the LR significance test to identify items with DIF. To date, however, $R^2\Delta$ has been used sparingly with LR in DIF research.

Jodoin (1999) proposed guidelines for interpreting $R^2\Delta$ (also see Gierl & McEwen, 1998). An item has negligible or A-level DIF when the chi-square test for model fit is not statistically significant or when $R^2\Delta<0.035$. An item has moderate or B-level DIF when the chi-square test is statistically significant and when $0.035 \leq R^2\Delta<0.070$. An item has large or C-level DIF when the chi-square test is statistically significant and when $R^2\Delta \geq 0.070$. These guidelines are applicable to both uniform and nonuniform DIF and were used to classify DIF items in this study.

Back-Translation

Back-translation is a popular and well-known judgmental method for evaluating the equivalence of two language forms (van de Vijver & Leung, 1997). In the basic design, the source language test is first translated into the target language, then back-translated into the source language by a different translator. The equivalence of the original source and target language forms is assessed by a reviewer or committee of reviewers who compare the original and back-translated source language forms for comparability in meaning (Brislin, 1970, 1986; Hambleton & Bollwark, 1991; Werner & Campbell, 1970).

The back-translation design has some notable advantages. For example, it enables the researcher who is not fluent in the target language to evaluate the quality of the test translation by comparing the original and back-translated source language forms. Researchers also agree that this method serves as a

general check on translation quality and that it can detect translation differences (Ellis, 1989; Hambleton, 1993; Hulin, Drasgow, & Komocar, 1982; van de Vijver & Leung, 1997). The back-translation design also has disadvantages. For example, the evaluation of test equivalence is conducted only in one language, and there is no assurance that the findings in the source language generalize to the target language because the source-to-target language translation is not directly assessed. This problem stems from the assumption that errors made during the original translation will not be made during the back-translation. However, this assumption may not hold in practice when, for instance, skilled translators make adjustments in the translation to ensure the items are equivalent even when the original source to target language items are different (Brislin, 1970; Hambleton & Bollwark, 1991; Hambleton & Kanjee, 1995). This outcome may also occur if the back-translator improves the quality of the test in situations where the original translation is poor (Hambleton, 1993). Finally, van de Vijver and Leung (1997) contend that the basic back-translation design may result in a literal translation at the expense of connotations, naturalness, and comprehensibility across languages, especially when translators know their work will be evaluated with back-translation.

Modified back-translation designs have been proposed to overcome some of the limitations in the basic design (Bracken & Barona, 1991; Brislin, 1986; Hambleton, 1993). In this study, another modification to the traditional back-translation design is presented and used. The proposed design is an attempt to overcome the limitations in the basic design as outlined above. The modified design is presented in Figure 1.

In the modified design, the source language test is translated into the target language by the test developer. Then the target language test is independently back-translated into the source language by two translators. Test equivalence is assessed by comparing the two back-translations and the original source language test. Differences between the two back-translations and the source language highlight potential translation problems. With two back-translators who work independently, individual differences between them will reduce the

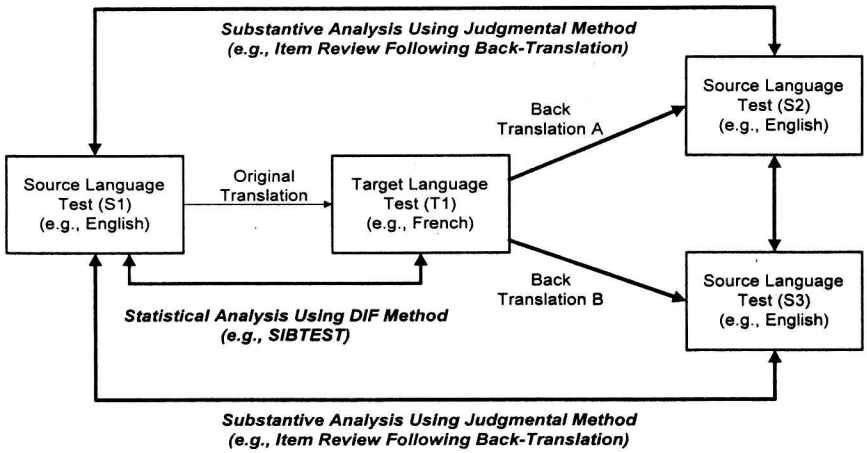

Figure 1. Framework for evaluating measurement equivalence for a test adapted from one source language (S1) to one target language (T1).

likelihood that each will make the same adjustments to ensure item equivalence when the original and source language items are in fact different. With two back-translators, it is also unlikely that both will make the same changes to improve the quality of a text in situations where the original translation is poor. Rather, different changes will be made resulting in inconsistencies between the two forms, thereby highlighting problems in the original translation. Finally, the modified design has a check; the researcher can evaluate back-translator consistency by comparing their translations with each other in addition to comparing their translations with the original source language.

Method

Student Sample and Achievement Tests

Data from 4,400 grade 6 students (2,200 English and 2,200 French Immersion) who wrote the 1997 administration of a provincial mathematics achievement test and 4400 grade 6 students (2,200 English and 2,200 French Immersion) who wrote the 1997 administration of a provincial social studies achievement test were used. Because Canada has two official languages, different language groups can be identified in many school districts. In this study the English-speaking examinees represent the dominant language group because most students receive instruction in this language at English-speaking schools. English-speaking students are tested in English. Alternatively, the French Immersion students are in programs where French is the language of instruction. Immersion programs are embedded in English-speaking schools. The Immersion program is designed for students whose first language is not French but who want to become functionally fluent in French and to develop an understanding and appreciation of French culture in addition to mastering English. Thus French Immersion students are linguistically distinct from English-speaking students. Immersion students are tested in French. The four samples were randomly selected from a database containing approximately 38,000 English- and 3,000 French-speaking grade 6 students for each test administration.

The mathematics test contained 50 multiple-choice items, and each item had four options. Test items were classified into five curricular content areas and two cognitive levels. The social studies test contained 49 multiple-choice items (the original test contained 50 items, but one item was dropped due to an obvious translation error), and each item had four options. Test items were classified into four curricular content areas and two cognitive levels. For both tests items were based on concepts, topics, and facts from the province-wide Program of Studies. The test score does not necessarily contribute to a student's final course grade, although teachers are encouraged to mark the tests and use the results for student grading.

All items were developed in English by a committee of item writers and a test developer 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 one French Immersion teacher and one Francophone teacher along with a bilingual test

developer. In this step the comparability of the English and French versions of the test was assessed by comparing the two forms. 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.

Statistical Analysis

Two of the three DIF statistics used in this study—Mantel-Haenszel and logistic regression—are based on the assumption that the test is unidimensional. To assess this assumption a confirmatory factor analysis was conducted. The indicator variables in the confirmatory factor analysis were created by summing the items in each curricular content area. A one-factor model was fitted to both the English and the French data for mathematics and social studies using LISREL 8.14 (Jöreskog & Sörbom, 1996). In addition, a multiple-sample analysis was conducted to evaluate the factor structure, factor loading, and error invariance across the English and French samples in the mathematics and social studies tests (Jöreskog, 1971). Three nested models were sequentially tested by equating the number of factors, factor loadings, and errors.

Confirmatory factor analytic models are assessed in part by using goodness-of-fit indices. Many different fit indices are available to the researcher, and yet despite their abundance few agree on which index provides the best answer to the question of model fit (Bollen & Long, 1993; McDonald & Marsh, 1990; Mulaik et al., 1989). As a result, three types of fit indices were used to assess each model. The first index is the chi-square statistic. Chi-square is used to determine if the restrictive hypothesis tested can be rejected. A model is considered to have acceptable fit if the difference between the variance-covariance matrix generated by the original data and by the hypothesized solution is small, yielding a nonsignificant chi-square. The chi-square statistic is dependent on sample size and often results in a statistically significant difference when large samples are used, even when fit appears good using other indices. Because chi-square is one of the most frequently used fit indices in a structural analysis, it was included (Elliott, 1994; Gierl & Mulvenon, 1995). The second index is the root mean square error of approximation (RMSEA). The RMSEA is intended to provide a measure of parsimony by assessing the discrepancy per degree of freedom in the model. In other words, RMSEA takes into account the number of free parameters required in order to achieve a given level of fit. Browne and Cudek (1993) suggest a RMSEA of 0.05 indicates a close fit of the model in relation to the degrees of freedom. They interpret a range of RMSEA values by stating, "we are also of the opinion that a value of 0.08 or less for the RMSEA would indicate a reasonable error of approximation and would not want to employ a model with a RMSEA greater that 0.1" (p. 144). The third index is the root mean square residual (RMR). RMR is an average of the

absolute residuals between the observed and the hypothesized covariances. A small RMR indicates good fit.

Next, DIF statistical analyses were conducted for each item from the English and French forms in the mathematics and social studies achievement tests using MH, SIBTEST, and LR. The item under consideration was included in forming the score groups for MH and LR, and no iterative purification was used in the DIF analyses. All test statistics were interpreted at an alpha-level of 0.05. Items that were flagged by at least two of the statistical procedures with B- or C-level were considered translation DIF items. This interpretation seems justified as B- and C-level items are typically scrutinized for potential bias in test reviews (Zieky, 1993).

Judgmental Analysis

Two translators independently back-translated the achievement tests from French to English. Both translators were certificated by the Association of Translators and Interpreters of Alberta, which is an association affiliated with both the Canadian Translators and Interpreters Council (CTIC) and the International Federation of Translators. To become certificated, translators must pass the national CTIC exam; to remain they must pass the CITC exam once every three years. Both translators have been accredited since 1990, and both have extensive experience translating business, industry, government, and educational texts from French to English. The translators were blind to the outcomes from the statistical analyses, and they had no contact with one another during this study.

Next three reviewers—ourselves—independently evaluated the comparability between the English source language tests and the back-translated tests. Each reviewer was given the original English form, the French translated form, and the two back-translated forms. The reviewers were explicitly asked to evaluate the comparability in meaning between the original English form and the two back-translated forms using a three-point rating scale: 1 = No Change in Meaning, 2 = Minor Change in Meaning, and 3 = Major Change in Meaning. The scale was used by the reviewers to identify items that were translated incorrectly.

Ratings were established with a four-step process. First, the three reviewers independently rated each item using the 3-point scale. All reviewers were blind to the outcomes from the statistical analyses at this step. Second, the three reviewers met to discuss and justify their ratings for each item. In this step the reviewers could change their ratings if they felt a change was justified based on the discussion. However, the reviewers were not required to change their ratings. Third, the reviewers compared the statistical outcomes with their judgmental ratings. The statistical outcomes were evaluated because, as Hambleton and Kanjee (1995) note, the ultimate criterion of item equivalence must come from an analysis of the examinees' responses. Thus the reviewers considered these outcomes as they evaluated the translation equivalence of each item. Again, changes were permitted but not required. Fourth, the three adjusted ratings were compared and a final rating was produced. The rules for creating the final rating were as follows: If at least two of the three ratings were

a 1, then the item was deemed equivalent; if at least two of the three ratings were either a 2 and/or 3, then the item was deemed not equivalent.

Results

Psychometric Characteristics of the Test Forms and Items

A summary of the observed psychometric characteristics on the mathematics and social studies tests for the English- and French-speaking examinees is presented in Table 1. Typically, the differences reported in Table 1 are tested for statistical significance between groups. However, the large samples used in this study resulted in many differences that were statistically but not practically significant. Hence statistical outcomes are not reported. Instead, some general trends are highlighted. First, the psychometric characteristics of the items were comparable between the English- and French-speaking examinees. The measures of internal consistency, difficulty, and discrimination were quite similar for both language groups in mathematics and social studies. Second, the mean for the French-speaking examinees was somewhat higher that the mean for the English-speaking examinees in mathematics; the order was reversed in social studies. However, for both tests the effect sizes associated with these mean differences were relatively small. Third, the standard deviations, skewness, and kurtosis were similar between the two groups for each test. Fourth, the number of words on the French forms was noticeably larger than the English forms, especially in social studies.

Factor Structure Within and Across Language Groups

Results from the confirmatory factor analysis supported the unidimensional assumption. The one-factor model provided excellent fit to the English and

Table 1
Psychometric Characteristics for the English and French Forms in Mathematics and Social Studies

	Mathematics		*Social Studies*	
Characteristic	*English*	*French*	*English*	*French*
No. of Examinees	2,200	2,200	2,200	2,200
No. of Items	50	50	49	49
No. of Words	2,713	3,066	3,354	4,157
Mean	35.44	37.12	32.67	31.75
SD	8.34	7.57	8.29	7.71
Skewness	−.49	−.65	−.44	−.34
Kurtosis	−.47	−.11	−.43	−.49
Internal Consistency[a]	.89	.87	.87	.84
Mean Item Difficulty	.71	.74	.67	.65
SD Item Difficulty	.15	.14	.12	.12
Range Item Difficulty	.26-.91	.22-.94	.39-.86	.39-.87
Mean Item Discrimination[b]	.48	.45	.43	.38
SD Item Discrimination	.12	.11	.11	.11
Range Item Discrimination	.05-.66	.09-.67	.15-.63	.17-.59

[a]Cronbach's alpha
[b]Biserial correlation

Table 2
Fit Indices for the One-Factor Model Across Content Areas as a Function of Language Group

Content Area	χ^2		*df*		*RMSEA*		*RMR*	
	English	*French*	*English*	*French*	*English*	*French*	*English*	*French*
Mathematics One-Factor Model	16.16*	21.75*	5	5	.032	.039	.034	.040
Social Studies One-Factor Model	2.08	1.99	2	2	.004	.000	.024	.025

*$p < 0.01$

French data on both the mathematics and social studies achievement tests, as shown in Table 2. Although the chi-square values were statistically significant for the English and French sample on the mathematics test, the RMSEA and the RMR were small, indicating good model fit.

Results from the multiple-sample analysis also suggested that the number of factors and factor loadings were invariant across language groups on the mathematics and social studies tests. The results of the multiple-sample analysis are provided in Table 3. The one-factor model was fitted separately for the English and French sample, and a chi-square statistic was computed to assess parameter invariance across the two groups. Three nested models were sequentially tested using this approach by equating the number of factors, factor loadings, and errors. For the mathematics test, Models 1 and 2 were not statistically different, whereas Models 2 and 3 were statistically different, indicating that the observed variables were not equally reliable across the two language groups. Despite this difference, the RMSEA and RMR for Models 1, 2, and 3 were small indicating strong model fit. A similar pattern of results occurred with the social studies test data as Models 1 and 2 were not statistically different but Models 2 and 3 were different. Again, however, the RMSEA and RMR were small for all three models, indicating good model fit. In short, when we take into account the sensitivity of chi-square to sample size and examine the RMSEA and RMR, there is strong evidence to suggest that the number of factors and the factor loadings are invariant across the English and French groups in mathematics and social studies.

Comparison of DIF Classification Using Three Statistical Procedures

Item classification produced by the three DIF procedures—Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and logistic regression (LR)—was compared across the mathematics and social studies tests. Classification consistency across the procedures is summarized in Table 4. In all comparisons that follow, items with a B- or C-level rating are considered DIF items, whereas those with an A-level rating are not.

For the mathematics test, MH and SIBTEST each identified six DIF items, whereas LR identified 10 items with uniform DIF and one item with nonuniform DIF. Four of the six items identified by MH (items 6, 8, 44, and 47) were

Table 3
Tests for Invariant Models Between English and French Examinees

Content Area	χ^2	*df*	*RMSEA*	*RMR*
Mathematics				
Model 1				
Equated Number of Factors	37.91*	10	.036	.040
Model 2				
Equated Number of Factors	38.59*	14	.028	.042
Equated Factor Loadings				
Model 3				
Equated Number of Factors	55.14*	19	.029	.051
Equated Factor Loadings				
Equated Errors				
Social Studies				
Model 1				
Equated Number of Factors	4.07	4	.003	.025
Model 2				
Equated Number of Factors	15.21	7	.023	.110
Equated Factor Loadings				
Model 3				
Equated Number of Factors	28.56*	11	.027	.150
Equated Factor Loadings				
Equated Errors				

Model Comparison	χ^2	*df*
Mathematics		
Model 1 vs. Model 2	.68	4
Model 2 vs. Model 3	16.55*	5
Social Studies		
Model 1 vs. Model 2	11.14	3
Model 2 vs. Model 3	13.35*	4

*$p<.01$.

identified by SIBTEST. All 6 items identified by MH (items 6, 8, 15, 40, 44, and 47) were identified by LR as displaying uniform DIF. Of the 6 items flagged by SIBTEST, 5 were identified by LR for uniform DIF (items 6, 8, 41, 44, and 47). The only item with nonuniform DIF, item 16, was identified by LR.

For the social studies test, many more items were flagged with DIF. MH identified 19 DIF items, SIBTEST identified 27 DIF items, and LR identified 27 items with uniform DIF and 2 items with nonuniform DIF. All 19 items identified by MH (items 2, 3, 5, 6, 11, 13, 17, 18, 22, 24, 25, 29, 30, 33, 40, 44, 45, 47, and 48) were also identified by SIBTEST and LR for uniform DIF. SIBTEST flagged the same 26 items as LR for uniform DIF (items 2, 3, 5, 6, 9, 11, 13, 16, 17, 18, 19, 22, 24, 25, 27, 29, 30, 33, 34, 35, 36, 40, 44, 45, 47, and 48). Only 2 items, 9 and 24, were identified with nonuniform DIF, and both were flagged by SIBTEST and LR. Item 24 was also flagged by MH.

Table 4
Classification Consistency Across the Three Differential Item Functioning Procedures as a Function of Test

	Mathematics				*Social Studies*			
	MH	*SIBTEST*	*LR-U*	*LR-N*	*MH*	*SIBTEST*	*LR-U*	*LR-N*
MH	6				19			
SIBTEST	4	6			19	27		
LR-U	6	5	10		19	26	27	
LR-N	0	0	1	1	1	2	2	2

Note. MH is Mantel-Haenszel; SIBTEST is the Simultaneous Item Bias Test; LR-U is logistic regression with uniform DIF, and LR-N is logistic regression with nonuniform DIF. The diagonal of each matrix indicates the total number of items flagged using each procedure and the off-diagonal indicates the number of matches across procedures.

Table 5
Correlation Coefficients Across the Four DIF Effect Size Measures as a Function of Test

	Mathematics				*Social Studies*			
	Δ–*MH*	$\hat{B}$	$R^2\Delta$–*U*	$R^2\Delta$–*N*	Δ–*MH*	$\hat{B}$	$R^2\Delta$–*U*	$R^2\Delta$–*N*
Δ–*MH*	—				—			
$\hat{B}$	–.96	—			–.99	—		
$R^2\Delta$–U	.91	.92	—		.93	.95	—	
$R^2\Delta$–*N*	–.03	–.01	–.00	—	.27	.29	.25	—

Note. Δ–*MH* is Delta-Mantel-Haenszel; $\hat{B}$ is the effect size measure in the Simultaneous Item Bias Test; $R^2\Delta$–*U* is R^2 change for the Group variable in logistic regression associated with uniform DIF; and $R^2\Delta$–*N* is R^2 change for the Total Score-by-Group Membership interaction term in logistic regression associated with nonuniform DIF. Both Δ–*MH* and $\hat{B}$ are directional tests. A positive Δ–*MH* indicates DIF in favor of the French examinees, whereas the opposite is true for $\hat{B}$. Because $R^2\Delta$ does not provide a directional test of DIF, the absolute value of Δ–*MH* and $\hat{B}$ are used when these effect size measures are correlated with the $R^2\Delta$.

Table 6
Results of the Mathematics Item Review by Three Raters

Item	*Statistical Flag (Rating)*	*Favors*	*Reviewer Rating*			*Overall Rating*
			Reviewer 1	*Reviewer 2*	*Reviewer 3*	
6	MH (C), SIBTEST (B), LR (C)	English	1	1	1	Equivalent
8	MH (B), SIBTEST (B), LR (C)	English	1	1	1	Equivalent
15	MH (B), LR (B)	French	1	1	1	Equivalent
40	MH (B), LR (C)	English	1	1	1	Equivalent
41	SIBTEST (B), LR (B)	English	2	1	1	Equivalent
44	MH (C), SIBTEST (C), LR (C)	French	3	3	3	Not Equivalent
47	MH (B), SIBTEST (C), LR (B)	French	3	3	3	Not Equivalent

Note. Ratings range from a score of 1 to 3 where 1 = No Change in Meaning, 2 = Minor Change in Meaning, 3 = Major Change in Meaning. In the overall rating column, equivalent means the English and French items are equivalent in meaning, whereas not equivalent means the English and French items are not equivalent in meaning.

Table 7
Results of the Social Studies Item Review by Three Raters

Item	Statistical Flag (Rating)	Favors	Reviewer Rating			Overall Rating
			Reviewer 1	Reviewer 2	Reviewer 3	
2	MH (C), SIBTEST (C), LR (C)	French	3	3	3	Not Equivalent
3	MH (B), SIBTEST (C), LR (C)	English	1	1	2	Equivalent
5	MH (C), SIBTEST (C), LR (C)	French	1	1	1	Equivalent
6	MH (C), SIBTEST (C), LR (C)	French	1	2	1	Equivalent
9	SIBTEST (C), LR (B)	English	1	1	1	Equivalent
11	MH (B), SIBTEST (C), LR (C)	English	3	3	3	Not Equivalent
13	MH (B), SIBTEST (C), LR (B)	French	1	1	1	Equivalent
16	MH (A), SIBTEST (B), LR (B)	French	1	1	1	Equivalent
17	MH (B), SIBTEST (C), LR (C)	English	3	3	2	Not Equivalent
18	MH (B), SIBTEST (C), LR (C)	French	1	1	1	Equivalent
19	SIBTEST (C), LR (C)	English	1	1	1	Equivalent
22	MH (B), SIBTEST (B), LR (B)	French	1	2	1	Equivalent
24	MH (C), SIBTEST (C), LR (C)	English	3	1	3	Not Equivalent
25	MH (C), SIBTEST (C), LR (C)	French	3	2	2	Not Equivalent
27	SIBTEST (C), LR (B)	French	3	2	3	Not Equivalent
29	MH (B), SIBTEST (B), LR (B)	French	1	1	1	Equivalent
30	MH (C), SIBTEST (C), LR (C)	English	3	1	1	Equivalent

Table 7 (continued)

Item	Statistical Flag (Rating)	Favors	Reviewer Rating			Overall Rating
			Reviewer 1	Reviewer 2	Reviewer 3	
33	MH (C), SIBTEST (C), LR (C)	English	1	1	1	Equivalent
34	SIBTEST (B), LR (C)	French	1	1	1	Equivalent
35	SIBTEST (B), LR (B)	French	1	1	1	Equivalent
36	SIBTEST (B), LR (B)	English	1	1	1	Equivalent
40	MH (C), SIBTEST (C), LR (C)	English	1	1	1	Equivalent
44	MH (B), SIBTEST (C), LR (B)	French	1	1	1	Equivalent
45	MH (B), SIBTEST (B), LR (B)	English	1	1	1	Equivalent
47	MH (C), SIBTEST (C), LR (C)	English	1	1	1	Equivalent
48	MH (B), SIBTEST (B), LR (B)	English	1	1	1	Equivalent

Note. Ratings range from a score of 1 to 3 where 1 = No Change in Meaning, 2 = Minor Change in Meaning, 3 = Major Change in Meaning. In the overall rating column, equivalent means the English and French items are equivalent in meaning, whereas not equivalent means the English and French items are not equivalent in meaning.

Relations Among DIF Effect Size Measures

Classification consistency can also be evaluated by examining the correlation between the DIF effect size measures. A strong correlation indicates a close relationship between the rankings of the items. As shown in Table 5, effect size measures were highly correlated across DIF procedures except the measure for nonuniform DIF. For the mathematics test the MH effect size measure Δ–*MH* was highly correlated with the SIBTEST effect size measure $\hat{B}$ at –0.96 and with the LR effect size measures for uniform DIF, $R^2\Delta$–*U*, at 0.91. Δ–*MH* and the nonuniform DIF measure R^2 Δ–*N* were correlated at –0.03. $\hat{B}$ and $R^2\Delta$–*U* were also highly correlated at 0.92, whereas $\hat{B}$ and $R^2\Delta$–*N* were correlated at –0.01. The correlation between the $R^2\Delta$ measures was zero. For the social studies test Δ–*MH* was highly correlated with $\hat{B}$ and $R^2\Delta$–*U* (–0.99 and 0.93 respectively), but not with ($R^2\Delta$–*N* (r=0.27). $\hat{B}$ and $R^2\Delta$–*U* were highly correlated at 0.95, whereas $\hat{B}$ and $R^2\Delta$–*N* were weakly correlated at 0.29. The $R^2\Delta$ measures also had a weak correlation at 0.25. The correlations between the uniform effect size measures and $R^2\Delta$–*N* were larger in social studies than in mathematics because the social studies test has more translation DIF items.

Results from Judgmental Analysis

Results for the judgmental analysis are reported in Tables 6 and 7 for the mathematics and social studies tests respectively. Only results for items flagged by at least two of the statistical procedures with B- or C-level DIF are presented and discussed due to space limitations. Consistency among the three reviewers was high for identifying the nonequivalent items. Two items on the mathematics tests and six items on the social studies tests were deemed not equivalent between the two language forms. One mathematics item judged to be nonequivalent in English and French is presented in Appendix A to illustrate translation differences. For this item the English form contained a 12-hour clock with AM and PM, whereas the French form used a 24-hour clock. Students were required to interpret an AM to PM time difference in this item, and this difference was more apparent when the 24-hour clock was used. Also, when the 24-hour clock is used the first two options on the French form are clearly incorrect. This item was identified both statistically and substantively using back-translation as being different between the two languages. This item also demonstrates how an important cultural difference can influence test development because in English the 12-hour clock is routinely used, whereas in French the 24-hour clock is frequently used.

The judgmental review also highlights two key points. First, the numbers of items identified statistically with DIF and identified substantively as not equivalent in translation were noticeably smaller in mathematics (7 and 2 respectively) compared with social studies (26 and 6 respectively), indicating that translation differences are more pronounced in social studies. Second, most of the differences between the two language forms in both content areas were *not* attributed to translation problems because most of the items with large statistical DIF flags—5 of 7 items in mathematics and 20 of 26 in social studies—showed no apparent translation differences. That is, factors and reasons other than translation differences are needed to account for the statistical outcomes when the English and French examinees are compared.

Conclusions and Discussion

The purpose of this study was to evaluate the equivalence of two translated tests using statistical and judgmental methods. Performance differences for a large random sample of English- and French-speaking examinees were compared on a mathematics and social studies achievement test. Items displaying DIF were flagged using three different statistical methods—Mantel-Haenszel, Simultaneous Item Bias Test, and logistic regression—and the substantive meaning of these flags was studied by comparing the back-translated form with the original English-version and with the statistical outcomes for each item on both tests.

Statistical Outcomes

Two main statistical findings were found. First, the classification results across procedures were relatively consistent, *but not identical*. The correlation between the effect size measures for uniform DIF were also strong, but not perfect. These results indicate that the three procedures produce relatively consistent item classification and effect size rankings, but some discrepancies were present. One explanation for these discrepancies may be found with the cut-points (i.e., A-, B-, and C-levels) used to identify DIF items. Are these cut-points comparable across procedures? Our results suggest that the MH cut-points are more conservative than either the SIBTEST or LR cut-points. In an attempt to establish a consistent and defensible pattern of DIF item classification, researchers may choose to use at least two procedures when screening items for translation DIF. Researchers and practitioners should also expect to identify fewer DIF items with MH compared with SIBTEST or LR.

Second, the LR effect size $R^2\Delta$–U was highly correlated with the effect size measures for MH and SIBTEST. Correlations between the nonuniform LR effect size measure $R^2\Delta$–N with MH and SIBTEST were negligible because the later measures were not designed to flag nonuniform DIF. Thus $R^2\Delta$ appears to be a reliable measure that provides useful information when used with the LR statistical analysis for quantifying the magnitude of uniform and nonuniform DIF. New guidelines for interpreting the LR and $R^2\Delta$ results for classifying items with DIF were proposed by Jodoin (1999), but much more research with LR and $R^2\Delta$ is needed to establish the validity of this procedure for identifying DIF. The first author is currently conducting simulation studies in this area.

Interpretability of DIF

The psychometric literature contains an "arsenal" of statistical approaches for identifying DIF (Angoff, 1993, p. 21). Despite the presence of these approaches, many researchers agree that items with DIF are difficult to interpret (Camilli & Shepard, 1994; Hambleton & Jones, 1994; O'Neill & McPeek, 1993; Scheuneman, 1987; van de Vijver, 1994). Camilli and Shepard reported that in their experience as many as half of the items with "large" DIF statistical values might not be interpretable. This finding may be attributable to statistical Type I error or to complex sources of item difficulty that cannot be interpreted using judgmental reviews. In an attempt to overcome this problem, multiple statistical procedures and a multifaceted judgmental review were used in the present study to identify items with translation differences. With these procedures, 2 of 7 and 6 of 26 DIF items in mathematics and social studies respectively were

judged to be nonequivalent across language forms due to differences introduced in the translation process. There were no apparent translation differences for the remaining items.

Translation differences were more pronounced in social studies, a language-rich content area. The sheer number of words on the social studies test (3,354 and 4,157 on the English and French forms respectively) compared with the mathematics test (2,713 and 3,066 on the English and French forms respectively) could lead to more translation differences. Student performance could also be affected. French-speaking examinees in social studies, for example, were required to read 803 more words compared with the English-speaking examinees, an increase of almost 24%, and this increased reading load could adversely affect test performance. The word count difference across language forms may account for some of the discrepancy between the statistical analysis and the judgmental review.

The discrepancy between statistical and judgmental results in social studies may also be attributed to inflated Type I error resulting from the use of an inadequate conditioning variable. DIF analyses require a conditioning variable that matches examinees in both language groups on the same construct of interest. When the number of flagged items becomes large, indicating many potentially problematic items, total test score (or a latent version of total test score as with the SIBTEST procedure) may not be a valid conditioning variable. Currently there is no solution to this problem when the number of DIF items is large (Sireci, 1997; Sireci, Xing, & Fitzgerald, 1999). As a first step, care must be taken to ensure the translation is accurate. Iterative purification can also be used by removing the DIF items from the total test score and repeating the analysis. However, this approach has been studied with only a relatively small number of DIF items (Clauser, Nungester, Mazor, & Ripkey, 1996). When a large number of items are flagged on a unidimensional test as in social studies, it is not clear which items to remove, as the purpose of the DIF analysis is to identify problematic items or how the construct and content representation on the test will be affected when a large number of DIF items are removed from the conditioning variable. This problem remains unsolved and must be addressed in future research.

Finally, researchers must also focus on *psychological* factors that produce DIF by studying items, actual student responses, and item-by-response interactions to identify and understand the sources of variability that produce large DIF results when two language groups are compared. There is general acceptance in the psychometric community that the psychology of test performance must be understood in order to construct, score, and validly interpret results from tests (Frederiksen, Mislevy, & Bejar, 1993; Gierl, 1997; Mislevy, 1996; Nichols, 1994; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1989). Despite this consensus, little is known about the cognitive processes actually used by examinees as they respond to test items in different language forms. To understand and interpret DIF better, it is necessary to study relations among cognition and task performance by examining *students' cognitive processes* as they respond to test items in different content areas and by creating cognitive models that will allow us to compare and contrast student performance. This research is essential because DIF statistics that are uninterpretable do not

provide practitioners with the information needed to make decisions about a test (e.g., which DIF items to drop and which items to keep when B- or C-level DIF is found) or about the psychological factors that produce DIF that should be considered when creating a test. In short, Camilli and Shepard (1994) correctly state that it is essential to "worry as much about how to interpret DIF as how to compute it" (p. 153). Researchers should address this concern by studying the psychological factors that produce translation differential item functioning.

Acknowledgements

This research was supported with funds awarded to the first two authors from the Social Sciences and Humanities Research Council of Canada (SSHRC), the Support for the Advancement of Scholarship Fund (SAS) at the University of Alberta, and the Social Science Research Fund (SSR), also at University of Alberta.

We would like to thank Ronald K. Hambleton, University of Massachusetts at Amherst, for his critique on an earlier version of this article.

References

Allalouf, A., & Sireci, S.G. (1998, April). *Detecting sources of DIF in translated verbal items.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Angoff, W. (1993). Perspective on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Erlbaum.

Bollen, K.A., & Long, J.S. (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Bracken, B.A., & Barona, A. (1991). State of the art procedures for translating, validating, and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119-132.

Brislin, R.W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185-216.

Brislin, R.W. (1986). The wording and translation of research instruments. In W.J. Lonner & J.W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Newbury Park, CA: Sage.

Browne, M.W., & Cudek, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Budgell, G.R., Raju, N.S., & Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309-321.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: 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.

Clauser, B.E., Nungester, R.J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33,* 202-214.

Douglas, J., Roussos, L., & Stout, W., (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33,* 465-484.

Elliott, P.R. (1994, April). *An overview of current practice in structural equation modeling.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ellis, B.B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74,* 912-921.

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.

Gierl, M.J., & Mulvenon, S. (1995, April). *Evaluating the application of fit indices to structural equation models in educational research: A review of the literature from 1990 through 1994.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

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

Hambleton, R.K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. *Bulletin of the International Testing Commission, 18,* 3-32.

Hambleton, R.K., & Jones, R.W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly, 18,* 23-36.

Hambleton, R.K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147-157.

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.

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818-825.

Jodoin, M. (1999). *Reducing type I error rates using an effect size measure with the logistic regression DIF procedure.* Unpublished master's thesis, University of Alberta.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-366). Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

McDonald, R.P., & Marsh, H.W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247-255.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C.D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430-445.

Narayanan, P., & Swaminathan, H. (1994). Performance of Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning. *Applied Psychological Measurement, 18,* 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20,* 257-274.

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., & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Erlbaum.

Pope, G.A. (1997). *Nonparametric item response modeling and gender differential item functioning analysis of the Eysenck personality questionnaire.* Unpublished master's thesis, University of Northern British Columbia.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-116.

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

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24,* 97-118.

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.

Sireci, S.G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16,* 12-19.

Sireci, S.G., Xing, D., & Fitzgerald, C. (1999, April). *Evaluating adapted tests across multiple language groups: Lessons learned from the IT industry.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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, Macmillan.

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

Thomas, D.R., & Zumbo, B.D. (1996, July). *Variable importance in regression and related analyses.* Paper presented at the annual meeting of the Psychometric Society, Banff.

van de Vijver, F.J.R. (1994). Item bias: Where psychology and methodology meet. In A. Bouvy, F.J.R. van de Vijver, P Boski, & P. Schmitz (Eds.), *Journeys into cross-cultural psychology* (pp. 111-126). Lisse, Netherlands: Swets & Zeitlinger.

van de Vijver, F, & Leung, K. (1997). *Methods and data-analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Werner, O., & Campbell, D.T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Naroll & R. Cohen (Eds.), *A handbook of method in cultural anthropology* (pp. 398-420). New York: Columbia University Press.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

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

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores.* Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D., & Thomas, D.R. (1996, October). *A measure of DIF effect size using logistic regression procedures.* Paper presented at the National Board of Medical Examiners, Philadelphia.

Appendix A

Items 47 on the English and French form of the Grade 6 Mathematics Achievement Test respectively

47. On the first day of filming, the crew arrived on the set at 5:20 A.M. They left the set at 8:15 P.M. How long did the crew spend on the set that day?

    A. 3 h 5 min
    B. 5 h 5 min
    C. 13 h 35 min
    D. 14 h 55 min

47. Le premier jour du tournage, l'équipe arrive au plateau de projection à 5 h 20 du matin. Elle quitte le plateau à 20 h 15. Combien de temps l'équipe est-ce que l'équipe passe sur le plateau le premier jour?

    A. 3 h 5 min
    B. 5 h 5 min
    C. 13 h 35 min
    D. 14 h 55 min