

**Validity of the Simultaneous Approach
to the
Development of Equivalent Achievement Tests
in English and French (Stage II)**

Jie Lin

W. Todd Rogers

Centre for Research in Applied Measurement and Evaluation

The University of Alberta

Abstract

The purpose of this paper was to evaluate the equivalence of achievement tests developed using the simultaneous translation approach. The process of item development and results of pilot testing were first reported by Rogers et al. (2003). The present paper begins with a description of the second item review and the selection of the items for the final Grade 9 Mathematics and Social Studies test forms. Performance differences of English and French examinees were then compared in terms of psychometric characteristics and differential item functioning, which were derived from LERTAP and SIBTEST respectively. The judgmental review by certified translators highlights two key points. First, the numbers of items identified as not equivalent in translation (meaning or form) were noticeably smaller in Mathematics (4 and 4) compared with Social Studies (8 and 23), indicating that translation differences are more pronounced in Social Studies. Second, the inter-rater reliability for Mathematics (0.82) was markedly higher than that for Social Studies (0.62), indicating better agreement in Social Studies than Mathematics. The analysis of field-testing data indicates that the English and French forms of Mathematics and Social Studies tests developed using the simultaneous test development approach are generally comparable in terms of psychometric characteristics, except that French examinees outperformed English examinees in both subjects. Social Studies yielded a higher percentage of DIF items than Mathematics, which is consistent with findings from previous research (Gierl et al., 1999). Based on the item review by certified translators, it is anticipated that translation differences contribute marginally to the DIF found. A preliminary analysis of teacher comments collected during field testing suggests that curriculum differences might have caused DIF in some items. It is hoped that teachers' questionnaires and students' think-aloud protocols to be conducted in the next stage will help to uncover the sources of DIF in these tests.

Introduction

The adaptation and translation of educational tests are becoming ever more important due to the marked increase in international, national and state/provincial testing in a time when an increasing number of students are studying in different languages. The Third International Mathematics and Science Study (TIMSS), for example, involved over 45 countries and 30 languages. Forty-one countries participated in the Program for International Student Assessment (PISA) in 2003. In Canada, the use of two official languages also necessitates the development of interchangeable tests in both English and French. In spite of the expectation that the original tests and the subsequent translations are equivalent in the constructs they measure, research has shown otherwise (e.g., Allalouf, Hambleton, & Sireci, 1999; Angoff & Cook, 1988; Budgell, Raju, & Quartetti, 1995; Ercikan 1998, 1999; Gierl, 2000; Gierl, Rogers, & Klinger, 1999; Hambleton, 1993; Sireci & Berberoğlu, 2000; Sireci, Fitzgerald, & Xing, 1998; Solano-Flores, Trumbull, & Nelson-Barber, 2002; Tanzer, 2005; van der Vijver & Tanzer, 1998). Hambleton (2005) presented a good example of poor test adaptation. In an international comparative study of reading proficiency, American students were asked to compare pairs of words and identify them as similar or different in meaning. For the pair “sanguine – pessimistic” only 54% (slightly above chance) of American students answered correctly. In the top-performing non-English-Speaking country, however, 98% of the students answered this question right. It was later discovered that there is no equivalent for “sanguine” in their language. Consequently the counterpart of “optimistic” was used, which in turn made the question a lot easier. This example provides a good illustration of how translation can affect the validity of a test in multilingual assessments.

An accepted and frequently applied procedure for translating tests is successive translation. Successive translation involves (a) test development in the source language by monolingual/monocultural test writers, (b) forward translation of the test into the target language by translators and review of the translated test by bilingual teachers or scholars, and (c) back translation of the translated test into the source language to monitor retention of the original meaning in the source language (Behling & Law, 2000; Hambleton & Bollwark, 1991). To the extent that the original and back translated versions of the test in the source language are similar, evidence is provided for the equivalence of the original and translated tests. The successive translation approach enables researchers who are not fluent in the target language to evaluate the

quality of translation by comparing the original and back-translated source language tests (Gierl et al., 1999; Hambleton, 2005). Researchers generally agree that the successive translation method provides an overall check on translation quality and can be used to detect translation differences (Ellis, 1989; Hambleton, 1993, 2005; Van de Vijer & Leung, 1997).

Despite the advantages, some concerns arise with the successive approach. First, Stanfield and Kahl (1998) contended that the differences between the original and translated tests might be due to problems with the back translation, not to problems with the forward translation. The back translation is just as likely to contain translation errors as is the forward translation. Basically, “one is left with two translations with no verifications of the quality of either” (Stanfield & Kahl, 1998, p. 6). On the other hand, in the process of back translation skilled translators may improve the test when the original translation is poor (Hambleton, 1993). Without a direct evaluation of the source-to-target translation, one can never be certain whether the discrepancies between the original and back-translated tests in the source language are attributed to problems in forward translation or back translation.

Further, the translated version of a test may not capture entirely the thinking associated with language and culture (Greenfield, 1999). Even when the original and translated tests are linguistically equivalent, they may not be necessarily psychologically equivalent (Brislin, 1980). Monolingual developers’ lack of competence in other languages or cultures may lead to ethnocentrism and linguistic or cultural specifics in the source test that make it almost impossible to create equally “good” test versions in the target language (Rogers, Gierl, Tardif, Lin & Rinaldi, 2003). That is, monolingual/monocultural test developers of the original tests are usually experts in the subject matter as well as in the source language and culture, but they may not be equally knowledgeable in the target languages/cultures. As a result, the forward translators may find it difficult, if not impossible, to create a test in the target language that is equivalent to the original test in terms of linguistic, cultural, and psychological perspectives. Successive translation may also result in literal translation at the expense of connotation, naturalness and comprehensibility across languages, especially when the translators are aware that there will be a back-translation (Stanfield & Kahl, 1998; Van de Vijer & Leung, 1997).

Hambleton (2005) concluded that although successive translation method can “identify problems in a test adaptation process, it would rarely provide sufficient amount of evidence to support the valid use of an adapted test” (p. 13). In response to the above concerns, the

concurrent/simultaneous test development approach was proposed (Solano-Flores et al., 2002; Tanzer, 2005). Solano-Flores et al. (2002) used the word “concurrent” for this test adaptation approach to stress the fact that “the two languages converge or interact throughout the entire process of assessment development” (p. 111). Tanzar (2005) made a plea for simultaneous development as an alternative to ensure cross-lingual/cross-cultural validity.

In simultaneous test development, the test is explicitly created for use in a multilingual/multicultural assessment. When bilingual tests are developed simultaneously, bilingual/bicultural test writers develop the source and target forms at the same time, so the two forms are equally open to modification in the process of test development. Therefore, language and culture specifics can be detected and removed at the early stages of test development, thereby reducing the risk of construct bias and maximizing linguistic and cultural decentering in both construct clarity and test item relevance and representativeness (Rogers et al., 2003; Solano-Flores et al., 2002; Tanzer, 2005). The potential advantage of simultaneous test development is ensuring that the quality of the test is equally good across languages.

To date only one study has been conducted to investigate the utility of simultaneous test development approach. Solano-Flores et al. (2002) documented their task development process and concluded that concurrent test development “allows assessment developers to generate high-quality assessments for linguistic minorities by supporting them to give deeper consideration to subtle language issues and culture as part of their discussion throughout the entire process of assessment development” (p. 127). More research is needed to determine whether the hypothesized advantages of the simultaneous test development approach are indeed tenable with reasonable effort and cost.

The purpose of the present study was to investigate the validity and utility of simultaneous approach to the development of equivalent achievement tests in French and English. The major objectives of this study were: a) to develop Grade 9 Mathematics and Social Studies tests in French and English employing the simultaneous approach; b) validate the tests produced; and c) evaluate the utility of the simultaneous approach in terms of cost-effectiveness and ease of implementation (Rogers et al., 2003). The present paper will address two questions: (a) how comparable are the two language versions of the tests constructed using simultaneous test development approach? and (b) Is there evidence for differential item performance for

English and French examinees on the Grade 9 achievement tests in Mathematics and Social Studies developed using the simultaneous approach to test development?

Overview of Research Design

The research design involved three stages. First, six bilingual item writers, three for Social Studies and three for Mathematics, were recruited to develop the initial French and English versions for each item at the same time. After the item writers reviewed each other's work, the surviving items were pilot tested and the results analyzed (Rogers, et al., 2003).

During the second stage a panel of six certified translators reviewed the items in the pilot tests for common meaning and form. Following this review, one Mathematics test (28 items) and one Social Studies test (40 items) were constructed in both languages. The four forms of tests were then administered to provincial Grade 9 samples stratified by region as part of the field-testing conducted by Learning Assessment Branch (in May 2004). At the same time the teachers of the classes included in the field-test completed a teacher questionnaires in which they commented on each item with respect to item clarity, relation to the learner outcomes, and curriculum coverage.

The student responses were scored and analyzed using the LERTAP item analysis computer program (Nelson, 2000). Psychometric characteristics of the tests and items of two language versions were compared and summarized. Differential item functioning (DIF) analyses were performed using SIBTEST (Shealy & Stout, 1993), which tests DIF hypotheses and quantifies the size of DIF by estimating a measure of the effect size ($\hat{\beta}_{uni}$) for each item.

The third stage involves explaining the DIF found. First, the comments made by the teachers will be examined to determine whether the identified DIF was attributable to curriculum or translation differences. Second, to be completed in May 2005, a sample of DIF items and non-DIF items will be used for think-aloud interviews. Protocol analysis (Ericsson & Simon, 1993) will be conducted to compare the thinking patterns of students in the two language groups.

The first stage of the study was reported in Rogers et al. (2003). The present paper will address the second stage.

Second Stage

Method and Results

As indicated earlier, the present paper includes a description of the procedures used at stage two and the corresponding results. The procedures used and the results for each are given together given the sequential nature of the research design.

Item Review by Certified Translators

Six accredited translators, who are members of the Association of Translators and Interpreters of Alberta (ATIA), were recruited to review the items retained from the first pilot study. They independently evaluated the 59 Mathematics items and 51 Social Studies items in terms of the comparability in meaning and form. A three-point scale (1 - 'different', 2 - 'similar', and 3 - 'identical') was used to assess the comparability of meaning. In terms of the comparability in words, phrases, verb tenses, or form of expression (i.e., idiom), a *Yes* or *No* format was used. Once an item was identified to be different in meaning or form, the reviewers were asked to justify their ratings in the questionnaire.

Inter-rater reliability. G-Theory (Brennan, 2001) was used to assess the inter-rater reliability. A 6 x k (rater-by-item) random effects design was used, where k = 59 for Mathematics and k = 51 for Social Studies. The inter-rater reliability was 0.82 for the Mathematics items and 0.62 for the Social Studies items. Given the greater dependency on words in Social Studies, and the greater time needed to construct and initially edit the Social Studies items, of vocabulary involved in Social Studies tests, however, the discrepancy between the inter-rater reliabilities is not surprising.

Comparability of items. The median rating for 4 of the 59 (6.8%) Mathematics items and 8 of 51 (15.7%) Social Studies items was less than or equal to 1.5, indicating that these items were not equivalent in meaning. Four (6.8%) Mathematics items were identified by three or more translators to be different in at least one of the following areas: words, phrases, verb tenses, or form of expressions. For Social Studies, 23 (45.1%) items were considered different in at least one of the four areas.

Revision and selection of items for field-testing. All the items identified to be different in meaning or form were examined and revised where possible. Three members of the research team completed the revisions. The nature of simultaneous test development was maintained throughout the revision process: an item revised in one language was then checked in the other

language to ensure comparability before moving to the next item. Two items in Mathematics and one in Social Studies that differed in meaning were deleted due to the difficulty in achieving equivalency between English and French versions. The remaining revisions were not extensive. For example, the stem of one item originally asked: “What are the two most important factors that contributed to the industrialization process in England?” The word “process” was deleted because it did not appear in the French version. The new stem became: “What are the two most important factors that contributed to the industrialization in England?”

Next, 28 items were selected for Mathematics and 40 for Social Studies. These numbers matched the numbers in the other pilot tests developed by Alberta Education. The final selection of items was based on the test specifications, with the items distributed proportionally in terms of content category (topics) and thinking levels (knowledge or skills). Items not in need of revision were selected first in each cell of the tables of specifications followed by the revised items. For Mathematics tests, two items that were revised due to non-equivalent meaning were included in the final forms. For Social Studies, seven items that were revised due to non-equivalent meaning were included in the final forms. The inclusion of these revised items was necessary to ensure coverage of the tables of specifications.

Differential Item Functioning (DIF)

Differential item functioning (DIF) analysis is a procedure used to identify items that function differently between different groups, and thus help monitor the validity and fairness of tests. It is based on the assumption that test takers who have similar knowledge (based on total test scores) should perform in similar ways on individual test questions regardless of their sex, race, or ethnicity. DIF occurs when an item is substantially more difficult for one group than for another group after the overall differences in knowledge of the subject tested are taken into account. Once the DIF items are detected statistically, there is a need for substantive interpretation to determine whether the items display bias or impact. *Item bias* is generally defined as invalidity or systematic error in how a test item measures a construct for the members of a particular group (Camilli & Shepard, 1994). *Item impact* refers to group discrepancy in item performance that reflects actual knowledge and experience differences on the construct of interest. If the item is biased, which unfairly favours one group of examinees over another, the item should be deleted or revised. If the item demonstrates impact, which reflects the actual

difference in knowledge between the groups, the item should be retained but further investigation may be necessary to explore why one group scored higher for this item.

To evaluate whether and to what extent biased test items existed in the two tests, DIF analyses were performed. An exploratory three-step approach (Camilli and Shepard, 1994; also see Roussos & Stout, 1996; Ramsey, 1993; Zieky, 1993) was used:

1. The simultaneous item bias test (SIBTEST) was used to find items for which there are unexpected differences in performance between two groups;
2. Each item that displayed DIF was examined in an attempt to identify the reason why the item was relatively more difficult for a particular group of examinees; and
3. An item was considered to be biased if it was established that the source of the unexpected or “extra” difficulty for one group was not relevant to what the test measures; otherwise the source of the DIF was undeterminable and in need of further research. (Camilli and Shepard, 1994, p. xiii)

SIBTEST. The simultaneous item bias test (SIBTEST), a nonparametric statistical method of assessing DIF in an item or bundle of items, is based on Shealy-Stout’s (1993) multidimensional model for DIF. The basic assumption is that multidimensionality produces DIF. SIBTEST detects bias by comparing the responses of examinees in the reference and focal groups that have been allocated to the same bins using their scores on a "matching subtest" (Stout & Roussos, 1995). The matching subtest is a subset of items that, ideally, are known to be unbiased.

SIBTEST was selected for use in this study mainly for three reasons. First, the number of students assessed in the field test was between 263 to 470 across the four language/subject test forms, thereby precluding the use of parametric procedures like the item response models that require larger numbers of students. Second, a number of studies have suggested that SIBTEST is more powerful and accurate in detecting DIF than other non-parametric procedures (e.g., Mantel-Haenszel) and parametric procedures like logistic regression procedures that are not dependent on large sample sizes (Bolt & Stout, 1996; Ercikan, Gierl, McCreith, Puhan, & Koh, 2002; Gierl, Jodoin, & Ackerman, 2000; Gierl, Rogers, & Klinger, 1999; Jiang & Stout, 1998). Third, SIBTEST uses a regression estimate of the true score as the matching variable, so that the examinees are matched on a latent rather than an observed score (Gierl, Rogers, & Klinger, 1999).

The amount of DIF in the studied item is reflected in the effect size estimate $\hat{\beta}_{uni}$, 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. The true scores are estimated using linear regression and then adjusted using a regression correction technique (Shealy and Stout, 1993). The weighted mean differences between the reference and focal groups on the studied item across the k subgroups is given by

$$\hat{\beta}_{uni} = \sum_{k=0}^k P_k d_k,$$

where P_k is the proportion of focal group examinees in subgroup k and d_k is the difference in the adjusted means on the studied item for the reference and focal groups, respectively, in each subgroup k . $\hat{\beta}_{uni}$ has a standard normal distribution with a mean of 0 and standard deviation of 1 under the null hypothesis of no DIF. The statistical hypothesis tested by SIBTEST is

$$H_0 : \beta_{uni} = 0$$

versus

$$H_0 : \beta_{uni} \neq 0$$

A statistically significant value of $\hat{\beta}_{uni}$ that is positive indicates DIF against the focal group and a negative value indicates DIF against the reference group.

Roussos and Stout (1996) provided general guidelines for interpreting the magnitude of item DIF: (a) negligible or A-level DIF: Null hypothesis is rejected and the absolute value of $\hat{\beta}_{uni} < 0.059$; (b) moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq$ the absolute value of $\hat{\beta}_{uni} < 0.088$; and (c) large or C-level DIF: Null hypothesis is rejected and the absolute value of $\hat{\beta}_{uni} \geq 0.088$.

Exclusion of two mathematics items. Prior to beginning the analysis it was noted that the agency that completed the field-testing altered the wording of some items in the French forms. In the case of Social Studies, the changes were in text material that preceded the questions related to that material. While the changes made were not correct, it was felt that they would not alter importance. However, this was not the case for Mathematics. Basically, English words were mistakenly used instead of French words. One of the two items is presented below:

L'entreprise A de location de voiture demande 32,00 \$ plus 1,10 \$/km tandis que l'entreprise B demande 37,00 \$ plus 0,90 \$/km. Quelle entreprise offre le meilleur prix pour une distance de 500 km et de combien?

- A. A by 95 \$
- B. A by 105 \$
- C. B by 95 \$
- D. B by 105 \$

The word “par” in French was replaced by “by” in English. In the second item, “et” was replaced by “and.” Therefore, these two items (Items 7 and 10) were deleted from all subsequent analyses.

Descriptive statistics. Descriptive statistics for the field test forms for Mathematics and for Social Studies are presented in Table 1. The mean test score for the French examinees was significantly ($p < 0.01$) higher than the mean for the English examinees on both the Mathematics and Social Studies tests. Next, the effect sizes (Glass & Hopkins, 1995, p. 290) were obtained by dividing the mean difference of the two groups by the standard deviation of the English examinees (i.e., the English group was considered the control group). For both tests the effect sizes associated with these mean differences were moderate: 0.41 for Mathematics and 0.33 for Social Studies. The standard deviations, skewness, and kurtosis were similar between the language groups for each test. Last, the internal consistencies were also comparable across groups, and are equivalent to 0.85 when the number of items is increased to the number of items included in the corresponding provincial test.

DIF: Mathematics tests. As shown in Table 3, two of the 26 mathematics items analyzed were identified with moderate DIF ($0.059 \leq \hat{\beta}_{uni} < 0.088$) and four items were identified with large DIF ($\hat{\beta}_{uni} \geq 0.088$). Three DIF items (one moderate and two large) favored English examinees, and three DIF items (one moderate and two large) favored French examinees. Among the four content categories covered in this test, *Patterns and Relations* has the highest percentage of DIF items (42.9%) while *Statistics and Probability* has none. For the three content categories that contain DIF items, the items favouring English and French examinees are roughly balanced, which suggests topic may not be a factor in the differential performance between the two language groups. As mentioned earlier, two items that initially differed in meaning were

included in the final forms. One of these revised items exhibited C-level DIF in favour of English examinees (in the area of *Patterns and Relations*).

DIF: Social studies tests. As shown in Table 4, 17 (42.5%) of the 40 items were identified with DIF: eight with moderate DIF and nine with large DIF. Ten (6 moderate and 4 large) of the 17 items favoured English examinees, and seven (2 moderate and 5 large) favored French examinees. Among the four content categories covered in this test, *The Former USSR* had the highest percentage of DIF items (75%) while *Quality of Life* contained 28.6% of DIF items. Interestingly, all the DIF items in *The Former USSR* and *Technology and Change* favoured English examinees. In contrast, all the DIF items in *Quality of Life* favoured French examinees. For *Economic Systems*, five of the eight DIF items favoured French examinees. There are two possible explanations for this phenomenon: one is ability differences and the other is curriculum differences. Further, it should be noted that two of the DIF items (one in *Economic Systems* and the other in *The Former USSR*) were among the seven items that were substantially revised following the review completed by the six ATIA reviewers. They both displayed B-level DIF in favour of English examinees.

Discussions and Conclusions

The purpose of this paper was to evaluate the equivalence of English and French versions of achievement tests constructed using the simultaneous test development approach. Prior to the field-testing, certified translators reviewed the Grade 9 Mathematics and Social Studies tests items that remained following the first test (Rogers et al., 2003) to determine if translation differences existed.

The judgmental review revealed two key points. First, the number of items identified as not equivalent in translation (meaning or form) was noticeably smaller in Mathematics (4 and 4) than in Social Studies (8 and 23), indicating that translation differences are more pronounced in Social Studies. Second, the inter-rater reliability for Mathematics (0.82) was markedly higher than that for Social Studies (0.62). The reviewers agreed more on the Mathematics items than on the Social Studies items, which corresponds to what was found during item development stage (see Rogers et al., 2003).

Descriptive statistics of the tests revealed that French examinees outperformed English examinees on both Mathematics and Social Studies tests. This corresponds to the trend observed

in Alberta provincial tests. Two possible explanations for the difference were advanced---the French examinees were from families with higher resources than the English examinees or the items in the two languages were not comparable in what they asked for (Rogers, et al. 2003).

The internal consistencies of the tests are analogous across the two language groups. When the number of items is increased to the number of items in the corresponding provincial test, the internal consistencies of the tests are estimated to be 0.85 using the Spearman-Brown prophecy formula (Spearman, 1904a, 1904b), which is comparable to the internal consistencies of Alberta provincial tests.

It is not surprising that Social Studies tests yielded a higher percentage of DIF items than Mathematics tests (42.5% vs. 23.1%). Gierl, et al. (1999), for example, found a similar pattern: 30.9% of DIF items in Alberta Grade 9 Social Studies tests and 22.4% of DIF items in the corresponding Mathematics tests (identified using SIBTEST). In general, Social Studies tests involve more vocabulary than Mathematics tests, and thus tend to produce more translation differences between the two language versions.

A more detailed study of DIF in the two tests revealed more differences in group performance. Among the six DIF items in the Mathematics tests, three favoured English examinees while the other three favoured French examinees. For the Social Studies tests, 10 of the 17 DIF items (58.8%) favoured English examinees, and seven (41.2%) favoured French examinees. That is to say, more DIF items were in favour of English examinees than French examinees in the Social Studies tests, while the numbers of item favouring English and French examinees were equal for the Mathematics tests.

An analysis of the distribution of DIF by content category (topic) demonstrated that content played different roles in the differential performance on the Mathematics tests and the Social Studies tests. For Mathematics, three of the four content categories contained DIF items, which were roughly balanced in terms of the language group they favoured. For Social Studies, however, topic did seem to make a difference in two groups' performance. Among the four content areas, all the items in *Former USSR* and *Technology and Change* favoured English examinees, while all the DIF items in *Quality of Life* and five of eight DIF items in *Economic Systems* favoured French examinees.

To sum up, the English and French forms of Mathematics and Social Studies tests developed using the simultaneous test development approach are generally comparable in terms

of psychometric characteristics, except that French examinees outperformed English examinees in both subjects. Social Studies yielded a higher percentage of DIF items than Mathematics, which is consistent with findings from previous research (Gierl et al., 1999). Based on the item review by certified translators, it is anticipated that translation differences contribute marginally to the DIF found. A preliminary analysis of teacher comments collected during field testing suggests that curriculum differences might have caused DIF in some items. Therefore, the next stage of the research will focus on further analysis of teacher comments as well as students' think-aloud protocols in an attempt to clarify whether or not the sources of DIF are attributable to translation errors, cultural issues, curriculum difference, and/or some other as yet unidentified reason.

Generally, the results of this study will have implications at the provincial, national, and international levels as government testing branches and private testing agencies increasingly face the need to conduct assessments in more than one language to more than one cultural group. Of particular concern is the situation in Canada where there is uncertainty about the equivalence of French and English versions of the same test, and the fact that, despite this uncertainty, comparisons are made among students and between the two language groups with their differing cultures. The findings of this study will contribute to a resolution of this uncertainty, and provide needed guidance for change to ensure the equivalence called for is in fact being achieved, thereby increasing the equity and fairness of our testing programs.

References

- Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. *Journal of educational measurement, 36*, 185-198.
- Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report 88-2)*. New York, NY: College Entrance Examination Board.
- Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks, CA: Sage.
- Bolt, D., & Stout, W. (1996). Differential item functioning: its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95
- Brennan, R. L. (2001). *Generalizability Theory*. Springer-Verlag: New York.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural research, 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. 137-162). 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.
- Douglas, J., Roussos, L., & Stout, W. (1996). Item bundle hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465-484.
- Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*, 912-920.
- Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*, 543-553.
- Ercikan, K. (April, 1999). *Translation DIF on TIMMS*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

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

Ericsson, K., & Simon, H. (1993). *Protocol analysis: Verbal report data*. Cambridge, MA: MIT Press.

Gierl, M. J. (2000). Construct equivalence of translated achievement tests. *Canadian Journal of Education*, 25, 280-296.

Gierl, M. J., & Khalig, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, 38(2), 164-187.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, SIBTEST, and Logistic Regression when the number of DIF items is large*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and judgment reviews to identify and interpret differential item functioning. *Alberta Journal of Educational Research*, XLV (4), 353-376.

Glass, G. V., & Hopkins, K. D. (1995). *Statistical Methods in Education and Psychology* (3rd ed.). Boston: Allyn and Bacon, Inc.

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. (2005). Issues, designs, and technical guidelines for adapting tests in to multiple languages and cultures. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

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., & 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.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Erlbaum

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.

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 and Behavioural Statistics, 23*, 291-322.

Jodoin, M. & Gierl, M. (2000, April). *Reducing type I error using an effect size measure with the logistic regression procedure for DIF detection*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Nelson, L. R. (2000). *Item analysis for tests and surveys using LERTAP 5*. Perth, Western Australia: Curtin University of Technology.

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. (2003). Differential Validity and Utility of Successive and Simultaneous Approaches to the Development of Equivalent Achievement Tests in French and English. *The Alberta Journal of Educational Research, 49*, 290-304.

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

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., & Berberoğlu, G. (2000). Using bilinguals to evaluate translated assessment questions. *Applied Measurement in Education, 13* (3), 229-248.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). Adapting credentialing examinations in international uses. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Solano-Flores, Trumbull, & Nelson-Barber (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, 2, 107-129.

Spearman, C. E. (1904a). 'General intelligence' objectively determined and measured. *American Journal of Psychology*, 5, 201-293.

Spearman, C. E. (1904b). Proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.

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

Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. Hambleton, P. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Erlbaum.

Tanzer, N. K., & Sim, C. Q. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptations. *European Journal of Psychological Assessment*, 15 (3), 258-269.

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

van de Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, 47 (4), 263-279.

Table 1

Item Review by Certified Translators

	Mathematics	Social Studies
Number of items	59	51
Inter-rater reliability	0.82	0.62
Number of items different in meaning ^a	4 (6.8%)	8 (15.7%)
Number of items different in form ^b	4 (6.8%)	23 (45.1%)

^a These items obtained 1.5 or lower on a three-point scale.

^b These items are identified to be different by three or more translators in at least one of the four categories: words, phrases, verb tenses or form of expressions.

Table 2

Psychometric Characteristics for Grade 9 Mathematics and Social Studies Achievement Tests

	Mathematics		Social Studies	
	English	French	English	French
No. of Examinees	469	345	470	263
No. of Items	26	26	40	40
Mean	13.14 ^a	15.04 ^a	23.74 ^b	25.75 ^b
Standard Deviation	4.64	4.56	6.11	6.12
Skewness	0.36	0.02	-0.14	-0.34
Kurtosis	-0.38	-0.47	-0.61	-0.67
Internal Consistency (Cronbach's alpha)	0.755	0.7545	0.79	0.81

^a $t = -5.81$, $df = 812$, $p < 0.01$, $\Delta = 0.41$

^b $t = -4.25$, $df = 731$, $p < 0.01$, $\Delta = 0.33$

Table 3

Distribution of DIF in the Mathematics Tests

Item Number	Content Category	DIF Level	Favouring Group
3	Number 7 ^a	B	French
8		C	English
9	Patterns and Relations 7	C	French
11		C	English
12		B	English
24	Shapes and Space 8	C	French
	Statistics and Probability		

^a Total number of test items in the particular content category.

Table 4

Distribution of DIF in the Social Studies Tests

Item Number	Content Category	DIF Level	Favouring Group
1	Technology and Change 10 ^b	B	English
3 ^a		C	English
4 ^a		C	English
9		B	English
11 ^a	Economic Systems 19	C	French
12 ^a		C	French
13		C	French
17 ^a		C	French
18 ^a		B	French
19		B	English
26 ^a		B	English
28 ^a		B	English
31		Quality of Life 7	B
36 ^a	C		French
37	The Former USSR 4	C	English
38		C	English
39		B	English

^a Items that are chained (sharing the same stimulus text).

^b Total number of test items in the particular content category.