

**Differential Validity and Utility of Successive and Simultaneous Approaches
to the
Development of Equivalent Achievement Tests in French and English**

W. Todd Rogers

Mark J. Gierl

Claudette Tardif

Jie Lin

University of Alberta

The question of how tests can be validly translated from one language to another is one of the most contentious questions in educational measurement today. In order to ensure fairness and equity (see Standard 9.7, American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Guideline A.1.7, *Principles for Fair Student Assessment Practices for Education in Canada*, 1993; Hambleton, 1994), translated tests are being used increasingly in educational testing to assess the knowledge and skills of students who speak different languages and/or come from different cultures. While the expectation is that these tests, initially written in the source language and then translated to the target language, are equivalent in the constructs they measure, researchers have found that this expectation is often not met (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, in press; van der Vijver & Tanzer, 1998). For example, Gierl et al. (1999) illustrated the problem with the item presented in Figure 1. The students who wrote the French language version of the test outperformed the students who wrote the English language version. However, this performance difference is likely due to the use of the 24-hour clock in the French version and the 12-hour clock in the English version.

An accepted and frequently used procedure for translating a test requires, first, that the test is developed by monolingual/monoculture test developers in the source language for use with students within the same monolingual/monoculture context. Second, one or a few translators *forward* translate the finished test into the target language. A panel of bilingual teachers and/or scholars then reviews the translated test, and changes are made as needed. Third, the translated test is then *back* translated into the source language to monitor retention of the original meaning in the source language (Behling & Law,

2000; Brislin, 1970, 1986; Hambleton & Bollwark, 1991). This method is often described as *successive translation*.

Even though a team of professional translators using the successive translation method may produce a multilingual version that is linguistically equivalent to the test written in the source language, these versions may not be psychologically equivalent. Further, a significant part of the socialization process in a culture is transmitted through language (Tanzer, in press; Valdés & Figueroa, 1994). What is needed to ensure the valid interpretation of an examinee's test score in a multilingual/multicultural assessment is construct or domain equivalence and instrument equivalence across the cultures corresponding to the source and target languages. The construct must possess domain clarity (Fitzpatrick, 1983). Additionally, the items included in the test must be relevant to psychological and cultural factors found in the source and target samples of students to be tested. Researchers tend to agree that the successive translation 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).

Despite this support, serious limitations of the successive approach remains (Greenfield, 1997; Tanzer, in press; Tanzer & Sim, 1999). For example, the final evaluation of test equivalence is conducted only in the source language, and there is no assurance that the findings in the source language generalize to the target language because the source-to-target translation is not directly evaluated. This problem stems from the assumption that errors made during the forward 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 & Bollward, 1991; Hambleton & Kanjee, 1995). This outcome may also occur if the back-translator improves the test in situations where the original translation is poor (Hambleton, 1993). Van de Vijver and Leung (1997) contend that the successive 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.

Further, while the monolingual/monoculture test developers of the source language test usually have the qualifications necessary to develop the test in the source language and awareness of the cultural and linguistic specifics as well the contextual aspects of their language and culture, they usually do not have nor do they need competence in other languages/cultures or formal training in cross-cultural psychology. These deficiencies can, unknowingly, lead to the ethnocentrism and linguistic/cultural specifics in the monolingual/monoculture test source that restrict equally "good" test versions in the

target language. It is very difficult, if not impossible, to adapt monolingual/monoculture developed tests to the same level of relevancy and representativeness in the target language/culture without modifying the test to such an extent that the level of instrument equivalence and, perhaps, construct equivalence is lowered to such a degree that cross language/cross culture comparisons are no longer valid. The problem is exasperated further if the translator or committee involved in the test adaptation does not have the full range of expertise need to produce equivalent source and target language tests (Greenfield, 1997; Tanzer, in press; Tanzer & Sim, 1999).

In response to these concerns, *simultaneous* translation has been suggested (Solano-Flores et al., 2002; Tanzer, in press). In simultaneous test development, the test is developed *explicitly* for use in a number of languages/cultures. Each language/cultural group and their speakers are simultaneously provided with the same opportunities to influence the development of the multi-language test forms. In this way, idiosyncrasies specific to a particular language (e.g., idioms unique to a language) or culture (e.g., social norms) can be detected and removed during the early stages of the test development thereby maximizing linguistic and cultural decentering in both construct clarity and test item relevancy and representativeness.

Unlike successive test development, simultaneous test development allows the influence and integration of information from committee members representing the different language and cultural groups to affect test development directly. With this approach, the risk of construct bias is reduced and the degree of linguistic and cultural decentering is enhanced because the source and target language versions are equally open to modification. However, the developmental effort and costs of tests constructed using simultaneous translation will likely be greater than the developmental effort and costs of tests constructed using the successive approach. Consequently, the successive approach is still frequently used and “will be frequently employed in the foreseeable future” (Tanzer & Sim, 1999, p. 262).

What is needed at this juncture is a series of well-controlled studies to determine whether or not the hypothesized advantages of the simultaneous test development approach over the successive test development approach are, indeed, tenable with reasonable effort and cost. Consequently, the purpose of this three-year research program is to evaluate the differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in the French and English languages. The major research objectives are: a) create a common domain of specifications to develop achievement tests in Mathematics and Social Studies at Grade 9; b) develop versions of each test in French and English employing the simultaneous and successive approaches to test development; c)

validate the tests produced; and d) compare the utility of the simultaneous and successive approaches in terms of cost-effectiveness and easy of implementation.

Method

Design

The three-year design to be followed to complete and verify the translations is presented in Figure 2. As shown, both versions are treated in the same way. Each revision of the simultaneously translated forms will be based on the results of the corresponding field tryout. All simultaneously, forward, and backward translated test forms will be administered at the end. The data from these administrations will be compared using differential item function analyses to identify items displaying DIF. A sample of these items together with a sample of items not displaying DIF will then be used for the think aloud interviews, the responses from which will be compared in an attempt to examine the comparability of solution strategies and thinking used by the students in both language groups.

The first three activities of the first year of the test development sequence shown in Figure 2 are reported in the present paper. The intent is to illustrate the simultaneous test development process and to present the first preliminary data to reveal how well the simultaneous translation process is working. The subsequent activities are to be completed during the next two years.

Subject Areas and Grade Level

French and English test forms were developed for Social Studies and Mathematics at the Grade 9 level. The Social Studies curriculum is more sensitive to differences in cultural values and preferences than the Mathematics curriculum. Further, Gierl et al., (1999) found that translation differences were more pronounced in Social Studies, a language rich content area, compared to Mathematics. By including both subjects, the findings in one content area will help illuminate the findings in the other content area. It is expected that there will be greater agreement among the tests developed using the forward and backward translation methods for Mathematics than for Social Studies and that this difference in agreement will be reduced, if not eliminated, in both subject areas when the tests are developed using the simultaneous translation method.

Grade 9 was selected as the grade level. Leighton, et al. (1999) found that students in Grade 9 were quite capable of verbalizing their thoughts and provide clear reasons for the answers to test questions. This skill will be critical for the examinees using the think aloud procedures to be completed in the third year of the present study.

Item Writers

Two 3-member item development teams developed the items for Mathematics and for Social Studies, respectively. They were all nominated by the staff at Alberta Learning. As shown in Table 1, there was one female on each team. French was the first language for one (item writer A) of the three item writers for Mathematics and the three item writers for Social Studies. One Mathematics item writer (A) used both French and English daily; the remaining two used English. Two Social Studies item writers (D and F) used both languages daily while the third used French. All of the item writers on both teams were experienced teachers, and, with one exception (F, Social Studies), they had taught the subject for which they developed items for at least five years. They were all presently teaching in French Immersion classes at the Grade 9 level.

The six item writers were confident about their French language competence, describing it as strong to very strong. In contrast, while the three item writers for Mathematics described their English competency as very strong, the three item writers for Social Studies were more tentative, with item writer E describing her competence as strong and item writers F and G indicating they were not sure.

The six item writers were also confident to very confident about their knowledge of the curriculum and the instructional procedures to follow. All teachers in Alberta follow a common curriculum. The teachers have program guides that delineate expected learning outcomes and contain suggested teaching approaches and reference materials.

While item writers A, B, and E were confident about their knowledge of shared meanings and cultural specifics of French and English and cross-culture psychology (Goodwin & Lee, 1994), item writers C, D, and F were not equally confident about each of these three aspects. Item writers C, D, and F were unsure about their knowledge of the cultural specifics in their second language. Item writer C was also unsure about his knowledge of cross-culture psychology.

Turning to their background in test development, the three item writers for Mathematics had not completed an educational assessment course while the three members for Social Studies had. One writer on each team (A and D) had served as a writer for the provincial achievement testing program. None had previous translation experience. Lastly, three item writers (B, E, and D) were confident about their level of knowledge about test development, while the remaining three were less sure.

Item Development

Construct Clarity

The “level of thinking-by-subject matter” table of specifications used in Alberta for the provincial achievement tests at Grade 9 in Mathematics and Social Studies were used to define the constructs to be assessed. To ensure construct clarity, the item writers (teachers) on each development team first reviewed the table of specifications used for their subject area and the final number of items that would be needed after pilot- and field-testing all of the items they developed. These tables are presented in Table 2 for Mathematics and Table 3 for Social Studies. A condensed version of the Taxonomy of Educational Objectives: Cognitive Domain (Bloom, 1954) was then reviewed since the level of thinking dimension in both tables of specifications is based on this taxonomy. The item writers quickly reached consensus on what was to be assessed given that they were currently teaching their content area, the close alignment between the table of specifications and the provincial curriculum guide for each of the subjects, and, as mentioned earlier, the requirement that they all follow the program of studies of the province.

Item Writing

Each team member was then provided with a set of guidelines for constructing multiple-choice items (Hopkins, 1998) and four common types of translation errors identified on previous provincial achievement tests by an 11-member committee of test translators, editors, developers, and analysts (Gierl & Khaliq, 2001). These documents were reviewed and discussed.

Following this discussion, the nature of the item-writing task was explained. The item writers were asked *to write one item at a time in both languages*. They were allowed to choose what language they would first write the item. They were told that they could not move to the next item until they had *a) written the item in the second language, b) ensured that the items in both languages meant the same in both languages and c) called for the same level of thinking by students who would respond to the item in French and by students who would respond to the item in English*.

The review and discussion of the curriculum documents and the taxonomy, guidelines for item construction, the types of translation errors, and the instructions were completed in a ½ day. Following this discussion, each team member developed approximately 30 items each over the next 2-½ days. Each day started at 9:00 and ended at 4:00 with one-hour lunch break and coffee breaks determined by the

teachers. During the item writing sessions, the item writers sought advice from each other where the meaning of a word in one language was not clear or how to express a particular phrase, sentence, or question that was not clear.

Following the development of the items, the first and third authors of this paper met separately with each item development team to review and discuss each item. One of the research team members was fluently bilingual and possessed strong knowledge of the shared meanings and cultural specifics of the French and English language and culture and cross-culture psychology. The second research team member possessed expertise in the area of measurement and evaluation. Each item was thoroughly discussed before moving to the next item. The discussions were centered on the comparability between both language versions and correctness of the writing within each language version. Care was taken to ensure the integrity of the simultaneous translation process; no one language dominated these sessions.

The members of the Mathematics team met on a second occasion to review the placement of the items along the level of thinking dimension in the two-way table of specifications. Review of the initial placements by each item writer revealed that the items were not always consistently placed. This inconsistency was attributable to differences in the way the three teachers instructed mathematics. A similar meeting was not required for Social Studies.

Item Development Results

Reactions to Simultaneous Translation

The item writers were asked to provide their views about and reactions to the simultaneous translation process they engaged in at the end of the third item-writing day. First, they were presented with the following statement:

Some people claim that one big advantage of the simultaneous approach is that it ensures maximum linguistic and cultural comparability in the definition of the construct and the test items designed to measure it.

They then were asked to indicate the degree (1 = strongly disagree, ..., 5 = strongly agree) to which they agreed with this statement with respect to linguistic and cultural comparability. The results were somewhat mixed. Two of the three members on each team either agreed or strongly agreed with the above statement with respect to linguistic comparability. The third Mathematics item writer (item writer

B) was not sure while the third Social Studies item writer (E) disagreed. With respect to cultural comparability, the three Mathematics item writers indicated they were unsure. In contrast, two of the Social Studies item writers agreed that the simultaneous translation approach led to cultural comparability while the third item writer (E) was not sure. The difference between the two teams with respect to cultural comparability is attributable to difference between the nature of Mathematics and Social Studies. The Mathematics item writers were not sure how the French and English cultures were differentially involved. In contrast, culture and the values within culture form an important part of Social Studies.

Frequency of changes. The next two questions posed to the item writers concerned the frequency with which they changed the item as first written when writing it in the second language. First, though, it should be noted that all but one item in Social Studies were first drafted in French. When asked why, the item writers indicated that they had just finished teaching and that the teaching was in French classes. They said that it was just natural for them to do so.

The Mathematics teachers made changes less frequently than did the Social Studies teachers. This is not an unexpected result given the fixed nature of Mathematics compared to Social Studies. All item writers considered the opportunity to make changes during the first item development stage an advantage of the simultaneous translation approach. Two reasons were provided. First the item writers commented that any weakness in an item showed up immediately instead of later in the translation process. Second, the item writers indicated that there was a lack of loss of meaning due to the immediacy of the translation or, as one of the teachers put it, “the essence and objectives [to which are questions are referenced] are fresh in our minds.” The discussions that took place during the item writing revolved around the meaning of a word in one language and the comparability of the meaning of the corresponding word in the other language.

Difficulty of simultaneous translation. The item writers were asked to indicate how difficult they found the task of simultaneously developing an item in both French and English before moving to the next item. A five-point Likert scale (1 = not difficult at all, ..., 5 = very difficult) was used for this purpose and they were asked to explain their rating. The three ratings for the Mathematics teachers were 1, 2, and 2; for ratings for the Social Studies teachers were 2, 4, and 3. One Mathematics item writer (C) noted that “the only difficulty was in finding the appropriate term in English.” Social Studies item writer E made a similar comment: “Translation in English was quite challenging at times, and brought me back at times to modify the French version.” Item writer F added: “It slowed down the item writing process which may be a good thing.”

Strengths and Weaknesses of Simultaneous Translation

The item writers were asked to identify what they saw as the strengths and weaknesses of the simultaneous translation process for item development. The following attributes were identified as strengths:

- a. efficiency and speed;
- b. reduced loss of meaning because one version is written immediately after the other;
- c. better assurance that the level of language in both forms is suitable and incidental vocabulary does not confuse the students;
- d. immediacy of the process;
- e. helps us to be as specific as we can be in both languages;
- f. done at the same time by the same person, thereby avoiding differences that come up when one person prepares an item in one language and a second person does the translation; and
- g. allows for continuous revision of each item.

All six item writers commented on the fairness of the process for the students. Moreover, at the end of three days, they had about 90 items purposely developed to be equivalent in French and English.

Turning to the weaknesses, Mathematics item writer B felt there was a tendency to translate literally to the detriment of linguistic integrity. Social Studies item writer F felt that the need to translate quotations and tabular information published by various external agencies (e.g., provincial government, Statistics Canada) in one language to the second language was a weakness. These two points are true of any translation process and are not specific to the simultaneous translation process. Other concerns more specific to the simultaneous process were:

- a. the need to keep in mind and maintain a sharp focus across both cultures;
- b. process requires teachers who are really comfortable with the curriculum in both languages;
and
- c. may not be the best job done in English.

Review and Revisions

The research team members noted several grammatical errors and awkward wordings in the English versions of the items for both the Mathematics and Social Studies. This was not unexpected given the observation that French was the first language used to construct the items for all but one item, English was the first language of four of the teachers, and all were teaching in French. Consequently, the item writers came together again to review and revise their work. To preserve the simultaneous approach, agreement was reached on the revisions to be made to the French and English forms of each item before moving to the next item. No attempt was made to have a word-for-word translation and awkwardness in language was to be avoided. The first, for English, and third, for French, authors of the present paper facilitated this process by asking questions and making suggestions. The final decision to move to the next item was made by the item writers and not the members of the research team.

Approximately four hours was required to complete the revisions for Mathematics. To complete the same task for Social Studies required approximately nine hours. The extra time required for Social Studies is attributable to the greater use of words in Social Studies than in Mathematics and the association between language and values. For both teams the greatest amount of time was spent on agreeing on the wording for the French form of the item than on the wording for the English form.

Two items were deleted from the pool of Mathematics items and four items were deleted from the pool of Social Studies items because the item writers agreed that the French and English forms were not and could not readily be made equivalent.

Placement of Items in the Table of Specifications

Following the review and revision process, the placement of the items in their respective table of specifications was reviewed. Of particular concern was the placement of the items according to the level of thinking required. Several Mathematics items that assessed similar thinking levels were placed at both thinking levels within the table of specifications. This was not the case for Social Studies.

Consequently, the Mathematics item writers met to review the placement of their items along the level of thinking dimension. Altogether, they made 25 changes. Five changes involved moving an item to a different topic (e.g., from numbers to patterns and relations). The remaining changes involved level of thinking classification: three items were reclassified at the higher level and 17 were reclassified at the lower level. The discussion and reassignments centered on mathematical procedures and whether they

were known and could be applied “automatically” or whether some conscious thought was required. If it was the former, the item was classified at the knowledge level; otherwise it was classified at the skill level (see Table 2).

Pilot Test

Following the last review the Mathematics item pool contained 87 items and the Social Studies item pool contained 86 items. All items were written and revised simultaneously in French and English. The item writers were of the opinion that the items in both languages would equally assess the topic and thinking level to which they were referenced. Further, all item writers agreed that the simultaneous translation approach taken to first develop and then revise the items is workable with a caution that more time should be allowed to develop the items, particularly for subject areas like Social Studies for which the language is rich and values play a role.

The decision was then made to pilot test the items to determine the item characteristics to be used to guide further revision. The intent was not to test the equivalency of the forms at this point given that the pilot tests would be conducted in March and not toward the end of the school year.

Pilot Test Forms

The research team members developed two pilot test forms in both languages for Mathematics and Social Studies. The items were grouped by thinking level within each topic area (see Tables 2 and 3), and they were placed in the same order in both pairs of forms. The initial draft of the Mathematics pilot test forms contained 35 items; the initial drafts of the Social Studies forms contained 39 items. These numbers were determined by the total class time available, 50 minutes, to administer the pilot forms.

The two sets of item writers met together to review each form. They examined each item one at a time and compared once more the French and English to ensure correctness of expression and meaning. The changes made included correcting the spelling and accents in French for both Mathematics and Social Studies. Four items were deleted from the Social Studies forms, three because of lack of clarity in both languages and the fourth because of the lack of a clear reproduction of what was initially a colored map. Lastly, the item writers examined the items in the pool not included in the pilot test forms and were asked if any of these items should replace an item in the pilot tests. No changes were made. The final numbers of items in the Mathematics and Social Studies forms were both 35.

Given the date of the pilot test, the teachers in the sample classes would not have covered all the material in the curriculum (see Tables 2 and 3). Additionally, although all teachers in the province must teach the same material, not all teachers follow the same sequence when teaching the subject area topics. Consequently the students in the different classes would be exposed to different topics. Therefore, the teachers of the sampled classes completed a form on which they indicated whether they had taught, were presently teaching, or still needed to teach each of the subject area topics.

Pilot Test Samples

Each of the six item writers agreed to administer the pilot test forms in their French-speaking classes. Therefore, to control for school effects, each item writer arranged to have the pilot forms administered in English-speaking classes in their schools. The forms were counter-balanced to control for any class effects.

Pilot Test Results

The total sample size for the Mathematics forms and the Social Studies forms are shown in Table 4¹. Although these samples sizes are small, the results of the item analysis, conducted using LERTAP (Nelson, 2000), and the opportunity to learn information provided by the teachers are sufficient to guide the next round of revisions. Items were classified into three classes. Class A contains items for which the item discrimination index (the uncorrected point-biserial) was at least 0.20 for both language groups. Class B includes items for which the discrimination index was at least 0.20 for one language group and the majority of teachers for the other language group indicated that the topic was either presently being taught or was to be taught and items where the discrimination was less than 0.20, but positive, for both groups and the topic was either presently being taught or was to be taught in both the French and English classes. For example, the statistics and probability topic had been taught in a greater number of English classes than in French classes. The point-biserial for five of the 10 items referenced to this topic was greater than 0.20 for the English students but less than 0.20 for the French students. Class C contains the remaining items. The distributions of the items by class across the topic areas for each subject area are presented in Table 5.

¹ Although the teachers were asked to tell the students to answer all questions and to do their bests, the mathematics teachers in one school advised their students either to answer the items they wish or to answer only the questions that were related to material they taught. The data for the students of these teachers was incomplete. Consequently the responses from this school were not included in the analysis.

Mathematics. Of the 70 Mathematics, 27 items were in Class A, 30 items were in Class B, and 13 items were in Class C (First panel, Table 5). The distribution of the items in Classes A and B across the cells of the Table of Specifications suggests that at this point it will be possible to construct an examination of 50 relevant and representative items. Inspection of the distributions of item difficulties for Class A and Class B items within each language group revealed that the distributions were essentially uniform. The corresponding means and standard deviations were 0.49 and 0.22 for French and 0.44 and 0.17 for English, Class A and 0.46 and 0.20 for French and 0.32 and 0.16 for English, Class B. The observations that the item means for both groups are lower than those typically found on the provincial tests is attributable to the time of year the pilot test was administered (March and not June). The observation that the items means for the French students exceeds the corresponding means of the English students is attributable to the fact that the French students were French Immersion students and that, as reported by teachers, these students tend to have high socio-economic status. The sample sizes were not large enough to control for ability and conduct differential item functioning analyses. Rather, the intent of the pilot study was to obtain preliminary information on the performance of the items. This information reveals that, given the number of items in Class A and Class B, the range of difficulty for both language groups, and the distribution of the items across the cells of the Table of Specifications, it will be possible to construct a Mathematics examination of 50 relevant and representative items which, when administered toward the end of the year, will yield means and standard deviations commensurate with end-of-year performance.

Social Studies. Thirty-one of the 69 Social Studies items were in Class A, 26 items were in Class B, and 12 items were in Class C (Second panel, Table 5). The distribution of the items in Classes A and B across the cells of the Table of Specifications suggests that, with the exception of Technology and Change, it will be possible to construct an examination of 55 relevant and representative items. As for Mathematics, the distributions of item difficulties for Class A and Class B items within each language group are essentially uniform. The corresponding means and standard deviations were 0.68 and 0.16 for French and 0.49 and 0.13 for English, Class A and 0.48 and 0.24 for French and 0.32 and 0.15 for English, Class B. As for Mathematics, the item means for both groups are lower than those typically found on the provincial tests due to the time of the year at which the pilot test was conducted. Further, the items means for the French students exceed, to a greater degree than in Mathematics, the corresponding means of the English students. This finding is again attributable to the fact that the French Immersion students tend to have high socio-economic status and the greater amount of reading in the Social Studies items. However, the intent of the pilot study was to obtain preliminary information on the performance of the items. This information again reveals that it will be possible to develop a Social

Studies examination of 55 relevant and representative items subject to construct of additional Technology and Change items.

Work to be completed

Item revision and selection of items for final forms. Each item writing team will use the results from the item analysis of the pilot test to select and edit the items for the final French and English forms in both subject areas. The Social Studies team members will construct at least 10 additional Technology and Change items. At all times, the nature of simultaneous translation will be preserved. The final numbers of items selected will correspond to the numbers of items shown in Tables 2 and 3.

Item reviews. An expert panel of five reviewers will be formed for each subject area. To become familiar with the items, each member will be asked to take the tests composed of the selected items in both languages. Following this, they will mark their own work. The review team will then be instructed to consider both conceptual equivalence (Hambleton & Bollwark, 1991; Matsumoto, 1994) and linguistic equivalence (Lonner, 1985; Marsella & Leong, 1995) for each of the translated forms with the original form. The scales used by Jeanrie and Bertrand (1999) will be used.. To assess whether or not the translated version of each item retained the sense of each item, the following scale will be used (p. 281):

Referring to the meaning of the original item, the meaning of the translated item is:

- 1) identical 2) rather similar 3) rather different 4) different

The following scale will be used to assess whether or not the translated items contained comparable words, verb tenses, and idioms:

As compared to the original item, this translated item:

- A. uses a perfectly equivalent language, in its form and meaning
- B. uses an equivalent language in its meaning only
- C. uses an equivalent language in its form only
- D. does not use an equivalent language

Based on the findings of these reviews, the item writers will then make final revision to the items retained.

Successive translations. The retained items will then be successively translated. Three new English to French (for the English forms) and French to English (for the French forms) translators will independently forward translate each item within each test version into the other language. The translators will then come together to review each translation and reach a

consensus. The final forward translated form in each language will then be independently back translated by a second team of translators. As for the forward translations, the second teams will meet together to review each translation and reach a consensus. A second review group of three translators will then judge the equivalence of the back translated forms following the same process used to review the equivalency of the simultaneously constructed items.

Test administration. The four forms (Fr_{original} , En_{original} , $Fr_{\text{back translation}}$, and $En_{\text{back translation}}$) for each subject area will be administered to provincial samples stratified by region in May 2004. The anticipated sample sizes are 500 for English only, 500 for French only, and 500 for bilingual students. The administration procedures will be the same as those used with the provincial achievement tests.

Analysis. The student responses to the items in each test version will be scored and analyzed using the LERTAP item analysis computer program (Nelson, 2000). Classical test score item statistics will be computed for both the correct answer and each distracter.

Differential item function analysis (DIF). SIBTEST (Shealy & Stout, 1993) will be used to look for the presence of DIF in the seven pairs of tests identified by the double-headed arrows in Figure 1. DIF may be due to item bias or item impact. Two approaches will be used to see which of these sources explains the DIF found. A second review team of three translators will first examine the conceptual and linguistic equivalence of the source version and the target language version produced by the first review team (see Year 2 Translations). The members of this committee will work independently and then come together to discuss and justify their ratings. Following this discussion, they will be presented with the statistical results. They will then work independently to check to see if they wish to change any of their ratings. Following this, they will assemble for a second time to discuss and justify the revised ratings and to reach a consensus. Items identified by both statistical procedures and consensus agreement will be tentatively identified as biased items.

To further clarify the source of DIF, samples of 16 students corresponding to the seven test pair comparisons will be asked to think aloud as they respond to items that display DIF attributed to bias, DIF items thought not to be due to bias and therefore, to impact, and items that do not display DIF. Up to eight items of each kind will be used for Mathematics and Social Studies, with the number dependent upon the number of items with moderate to large DIF.

Students will be asked to think aloud while formulating their responses to each item. When necessary, they will be asked to clarify their responses and/or provide reasons for their answers after they completed an item. A standard set of non-directional probes will be used for this purpose (Ericsson & Simon, 1993). Each interview will be audiotaped for later transcription and analysis.

The tapes will be coded for the cognitive strategies such as use of declarative or procedural knowledge, level and organization of thinking, influence of culture, and use of meta-cognitive skills. The results from these analyses will help clarify any differences due to the test development process or to translation. This work will be completed by members of the research team and the graduate students research assistants who will take part in all aspects of this research study.

Implications

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 and assessment in more than one language to more than one cultural group. Rapid changes in economic, social, and education policy during the last decade have led to increasing demands to administer common tests in more than one language and to more than one culture. Of particular concern is the situation in Canada where we see common tests necessarily being administered in French and English in the national literacy, numeracy, and science assessments and in most provincial assessments. 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. This uncertainty has given rise to principles and standards drawing our attention to the need for construct equivalence across forms so that the interpretations made are valid and not open to misinterpretation. 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.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- 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 & Law, 2000
- 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.
- 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.
- Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement*, *7*, 3-13.
- 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., 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.
- Goodwin & Lee, (1994). Taboo topics among Chinese and English friends. A cross-cultural comparison. *Journal of Cross-Cultural Psychology*, *24*, 325-338.
- 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., & 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.

- 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.
- 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.
- Leighton, J. P., Rogers, W. T., & Maguire, T. O. (1999). Assessment of student problem solving on ill-defined tasks. *Alberta Journal of Educational Research, XLV* (4), 408-426.
- Lonner, W. J. (1985). Issues in testing and assessment in cross-cultural counseling. *The Counseling Psychologist, 13*, 599-614.
- Marsella, A. J., & Leong, F. T. L. (1995). Cross-cultural issues in personality and career assessment. *Journal of Career Assessment, 3*, 202-218.
- Matsumoto, D. (1994). *Culture influences on research methods and statistics*. Pacific Grove, CA: Brooks/Cole.
- 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.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (Third edition; pp. 13-103). New York, NY: American Council on Education and Macmillan Publishing Company.
- Nelson, L. R. (2000). *Item analysis for tests and surveys using LERTAP 5*. Perth, Western Australia: Curtin University of Technology.
- Principles for Fair Student Assessment Practices for Education in Canada*. (1993). Edmonton, AB: Joint Advisory Committee.
- 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.
- 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 Education 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
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.
- Tanzer, N. K. (in press). 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*. Hillsdale, 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.
- Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.
- 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 3
Background of Item Writers

Characteristic	Item Writer:	Mathematics			Social Studies		
		A	B	C	D	E	F
Gender		F	M	M	F	M	M
First Language		F	E	E	F	F	F
Language used Daily		F&E	E	E	F	F&E	F&E
Years of Teaching Experience		23	7	7	15	15	13
Years of Teaching Mathematics/Social Studies		7	5	5	8	9	1
Language used to Teach		F	F	F	F	F&E	F
Language Competence ^a							
in French		5	4	4	5	5	5
in English		5	5	5	4	3	3
Knowledge and Understanding of ^a							
Curriculum		4	5	4	5	4	4
Instructional procedures		5	4	4	5	4	4
Knowledge of ^a							
Culture specifics of French		5	4	3	5	5	5
Culture specifics of English		5	5	5	4	3	3
Cross-culture psychology		4	4	3	4	4	4
Test Development Background							
Completed an educational assessment course		No	No	No	Yes	Yes	Yes
Item writer for provincial testing program		Yes	No	No	Yes	No	No
Language used		Eng			Fr		
Previous translation experience		No	No	No	No	No	No
Knowledge of test development ^a		3	4	3	4	4	3

^aSelf-ratings of knowledge (1 = very weak, ..., 5 = very strong)

Table 2					
<i>Table of Specifications: Mathematics</i>					
					Topic
		Numbers	Patterns & Relations	Shape & Space	Statistics & Probability
Knowledge	<ul style="list-style-type: none"> recall facts, concepts, and terminology know procedures for algorithms and computations, and for using formulas know procedures for constructions, conversions, and order of operations know mental computation and estimation strategies know how to use calculators and computers 	4	4	5	3
Skills	<ul style="list-style-type: none"> apply basic mathematical concepts in familiar and unfamiliar situations demonstrate relationships among number systems, operations, number forms, and concrete, pictorial, and symbolic representations demonstrate and apply relationships within equations and formulas demonstrate and apply relationships among geometric forms in a variety of situations demonstrate relationships between numbers and geometric forms use a variety of strategies to solve problems apply data management skills to solve problems judge the reasonableness of a solution 	9	11	9	5

Source: Learner Assessment Branch, August 2001

Table 3					
<i>Table of Specifications: Social Studies</i>					
			Topic		
		Technology & Change	Economic Systems	Quality of Life in Different Economic Systems	The Former USSR
		Industrialization Technology	Market, Mixed, and Centrally Planned Economy		Geography & Economic Change
Knowledge	<ul style="list-style-type: none"> understands generalizations, concepts, related concepts, terms and facts 	9	9	2	2
Skills	<ul style="list-style-type: none"> locating interpreting organizing analyzing synthesizing evaluating 	12	12	6	3

Source: Learner Assessment Branch, August 2001

Table 4
Pilot Test Sample Sizes

Form	Content Area			
	Mathematics ¹		Social Studies	
	French	English	French	English
1	26	36	43	50
2	28	38	44	53

¹ Although the teachers were asked to tell the students to answer all questions and to do their bests, the mathematics teachers in one school advised their students either to answer the items they wish or to answer only the questions that were related to material they taught. The data for the students of these teachers was incomplete. Consequently the responses from this school were not included in the analysis.

Table 5
Distribution of Items by Class

		Topic								Total
		Number		Patt &	Rel.	Sha &	Space	Prob &	Stats.	
Level of Thinking:	Item Class ^a	K	S	K	S	K	S	K	S	
	A	2	8	2	6	3	2	1	3	27
	B	3	1	2	8	3	9	1	3	30
	C	0	2	0	4	1	3	3	0	13

		Topic								Total
		Tech &	Change	Eco	System	Qual of	Life	Former	USSR	
Level of Thinking:	Item Class	K	S	K	S	K	S	K	S	
	A	3	6	7	9	0	3	2	1	31
	B	3	1	6	5	3	6	0	2	26
	C	1	0	2	3	0	4	1	1	12

^a Class A: value item of the discrimination coefficient (the uncorrected point-biserial) is at least 0.20 for both language groups.

Class B: value of the discrimination index is at least 0.20 for one language group and the majority of teachers for the other language group indicated that the topic was either presently being taught or was to be taught and items where the discrimination was less than 0.20, but positive, for both groups and the topic was either presently being taught or was to be taught in both the French and English classes.

Class C: remaining items.

Figure 1. Item 47 on the English and French form of the Grade 6 Mathematics Achievement Test.
Figure 2. Translation Design

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

