

Validity of the Simultaneous Approach to the Development of Equivalent
Achievement tests in English and French (Stage III)

Jie Lin

W. Todd Rogers

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Paper to be presented at the annual meeting of NCME

San Francisco, April 10, 2006

Abstract

The present study is part of a research project designed to investigate the validity and utility of the simultaneous approach to the development of equivalent achievement tests in French and English. To address this purpose, three types of evidence were collected: judgmental review by certified translators, statistical analysis of the performance data, and think-aloud interviews with students. The first stage involved item development and pilot testing (Rogers et al., 2003), and the second stage involved item review, dimensionality assessment, and DIF analysis (Lin & Rogers, 2005). The purpose of the present paper was to describe the third stage which involved explaining the DIF found. All the DIF items and some non-DIF items were included in the question sets. Twenty-four English-speaking students and 40 French Immersion students were interviewed, and both concurrent and retrospective verbal reports were collected. In addition, French Immersion students were also asked to evaluate the comparability of the two language versions. Protocol analysis (Ericsson & Simon, 1993) was conducted to compare the patterns in which students from the two language groups understood the items.

For mathematics, no supporting evidence was identified to have caused the DIF found in four of the five items. The DIF in one item appears to be attributed to the interaction of heavy loads of information and inadequate proficiency in French for French Immersion students. For the 17 DIF items in social studies, no supporting evidence for DIF was found in nine items. For the other eight items, the DIF could be attributed to either differential familiarity with key words/concepts or differential difficulty of stimulus texts, which was caused mostly likely by the inadequate proficiency in French on the part of French Immersion students. Discussions and recommendations about the use of simultaneous approach were included.

Introduction

The adaptation of educational tests is 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 necessitates the development of interchangeable tests in both English and French. In spite of the expectation that the original tests and the subsequent adaptations are equivalent in terms of the constructs they measure, research has shown otherwise (e.g., Allalouf, Hambleton, & Sireci, 1999; 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 de Vijver & Tanzer, 1997).

The two most popular designs for adapting tests are forward translation and back translation. With the forward-translation design, one or more translators translate the test from the source language to the target language. Then the equivalence of the two versions of the test is checked by another group of translators. Revisions can then be made based on the recommendations of the second group of translators (Geisinger, 1994; Hambleton, 2005). The main advantage of the forward-translation design is that direct judgments can be made about the equivalence between the two versions of the test (Hambleton, 2005). Forward translation also involves less time compared with back translation. However, the weaknesses of the forward-translation design reside mainly in (a) the high level of inference that must be made by the

translators about the equivalence of the two version of the test, and (b) the inability of the monolingual test developers and researchers to judge test equivalence themselves (Hambleton, 2005). Although it may continue to be one of the most frequently used methods for test adaptation, “direct translation from the source language to the target language has been repudiated as an unreliable method for achieving language equivalence” (Brislin, 1970; Werner & Campell, 1970 in Erkut, Alarcon, Gracia Coll, Tropp, & Vazquez Garcia, 1999, p. 208).

The back-translation design is the best known and frequently applied procedure for adapting tests. A back-translation design involves (a) forward translation of the test into the target language by translators and review of the translated test by bilingual teachers or scholars, and (b) 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 back-translation design 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 back-translation design provides an overall check on adaptation quality and can be used to detect adaptation differences (Ellis, 1989; Hambleton, 1993, 2005; Van de Vijer & Leung, 1997).

Despite the advantages, concerns have been raised with the back-translation design. First, Stanfield and Kahl (1998) contended that the differences between the original and adapted tests might be due to problems with the back translation and 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 verification of the quality of

either” (Stanfield & Kahl, 1998, p. 6). Further, 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 the forward translation or the back translation. Second, the back-translation design may 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). Above all, a common weakness of both forward- and back-translation designs is that 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).

In response to the above concerns, Tanzer (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. Consequently, the multiple language 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 (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.

Purpose

The main purpose of this research project was to investigate the validity and utility of the simultaneous approach to the development of equivalent achievement tests in French and English. The major objectives of the large-scale project were to a) 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, ease of implementation, and quality of tests with regard to degree of biasedness and presence of measurement error (Rogers et al., 2003).

The first stage of the research project was reported in Rogers et al. (2003). In this stage, six bilingual item writers, three for mathematics and three for social studies, were recruited to develop the initial French and English versions for each item at the same time. They wrote the item in one language and then immediately translated it into the second language. They were not allowed to move on to the next item until they had made sure that the items in both languages meant the same and called for the same level of thinking for the target students. After the item writers reviewed each other's work, the retained items were pilot tested. The pilot test results revealed that the French-speaking examinees outperformed the English-speaking examinees in both mathematics and social studies. What was not clear is why this difference occurred. Possible reasons include non-equivalence of the tests constructed in the two languages, the presence of socio-economic differences between the two language groups (i.e., real differences in ability), or a combination of both (Rogers et al., 2003).

The second stage was reported in Lin and Rogers (2005). During this stage, the item writers met to revise the items in the pilot tests based on the item analysis of the pilot test results. Then a panel of five certified translators reviewed the retained items from the revision for

comparability in meaning and word use. The great majority of items in mathematics (93.1%) and social studies (94.1%) were considered identical or similar in meaning across the English and French versions. As expected, very few language differences were found in mathematics (5.8%), while more differences were identified in social studies (16.8%). Given the language-rich nature of social studies tests, however, item comparability was satisfactory. Revisions were then made based on the comments of the translators. One mathematics test (28 items) and one social studies test (40 items) were then constructed in both languages.

The four test forms were administered to a sample of Grade 9 students as part of the field-testing conducted by Alberta Education. The student responses from the field test were scored and item analyses were performed using the LERTAP item analysis computer program (Nelson, 2000). The psychometric characteristics of the English and French versions of the tests were generally comparable, with the one exception being that the French sample outperformed the English sample in both mathematics and social studies. The factor structures of the tests were then examined using NOHARM (Fraser, 1988) and LISREL 8.14 (Jöreskog & Sörbom, 1996). The results from the non-linear factor analysis and multi-group confirmatory factor analysis suggested that both the mathematics and social studies tests were unidimensional with comparable factor loadings and error variances across the two language versions. Finally, Differential Item Functioning (DIF) analyses conducted using SIBTEST (Shealy & Stout, 1993) indicated that 5 out of 26 (19.2%) mathematics items and 17 of 40 (42.5%) social studies items displayed DIF.

The present paper describes the third stage of the project, which involved explaining the DIF found. A sample of DIF items and non-DIF items were used for think-aloud interviews.

Protocol analysis (Ericsson & Simon, 1993) was then conducted to compare the patterns in which students from the two language groups understood the items.

Method

The purpose of the think-aloud interviews was to gain further understanding of the nature of the DIF found. An item was considered biased if it was established that the source of the extra difficulty for one group was not relevant to what the test measures (e.g., adaptation differences); otherwise the source of the DIF was considered to be undeterminable (e.g., real ability differences, curricular differences, or cultural differences) and in need of further research (Camilli & Shepard, 1994). In the case of this study, since the English-speaking students and French Immersion students attended the same schools and used the same curriculum, curricular differences were assumed to be minimal, if any. Similarly, as the English-speaking students and French Immersion students all had English as the first language and resided in English communities outside schools, the cultural differences were also expected to be negligible. Therefore, other than adaptation biases, the most likely factor contributing to DIF in this study was ability difference, or in other words, impact.

Since monolingual English-speaking students and bilingual French Immersion students were the major intended populations for the achievement tests developed in this study, a sample of students in each of these groups were interviewed for comparison of their understanding of the test items in English and French. Francophone students were not interviewed at this time mainly for three reasons. First, Francophone students represent only 10 percent of French-speaking Grade 9 population in Alberta. Second, given the small number of Francophone students and their scattered distribution across the province, it was not possible to draw a sample of Francophone students that could be interviewed given the limited resources of the present

study. Third, given the distinctiveness of the Francophone (French as the first language) population, the decision was made not to include individual Francophone students in the sample of French Immersion students (French as a second language).

Interview Sample

A sample of Grade 9 English-speaking students and a sample of Grade 9 French Immersion students were recruited from public schools in a major metropolitan area in Alberta that enrolled both English only and French Immersion students. It was made clear to the principals and teachers who assisted with the sampling that the students selected should be highly verbal in English or French and be able to think aloud while they solved the problems. Students were over-sampled to allow for non-response, denial, and absence. A total number of 200 students (80 English and 120 French Immersion) were selected to take information letters and consent forms to their parents, who then indicated whether or not they gave consent for their children to be interviewed on an individual basis. All together 31 consent forms for English-speaking students and 44 consent forms for French Immersion students were returned. In each case, the parents gave consent for the children to be interviewed. Sixty-four students--24 English and 40 French Immersion—were interviewed, with an equal number of males and females.

Instruments

The instruments used in the interviews included a mathematics or social studies assessment question set, a set of structured interview guidelines (see Appendix A), and a structured observation sheet (see Appendix B).

Assessment question sets. The assessment question set for mathematics contained five items that displayed DIF and four items that did not display DIF. The corresponding numbers for the social studies question set were 17 DIF and 5 non-DIF items. All the DIF items identified by

SIBTEST were included; the non-DIF items were included to mask the DIF items. The non-DIF items were selected to ensure their comparability in terms of difficulty, topic as well as format to the DIF items in each question set. The number of items administered in each interview was such that each interview could be completed in one class period. To meet the requirement, the nine mathematics items were placed in one form in English and in two forms (one with five items and one with four items) in French. For social studies, the 22 items were divided into two forms (each with 11 items) in English, and three forms (two with seven items each and one with eight items) in French. The smaller number of items in the French forms allowed time for the French Immersion students to evaluate the comparability of the English version and French version of the items in the interview question set.

Interviewing guidelines. The think-aloud protocol contained a set of questions that the interviewers posed to participants upon completion of each item. Drawing on Ercikan et al. (2004), the following four questions were used with all students:

1. Are there any words that you do not know in this question?
2. In your own words, could you tell me what you believe this question is asking?
3. Did you find any parts of the question confusing? If so,
 - i. What parts did you find confusing?
 - ii. Why are they confusing?
4. Did you find any parts of the question helpful in solving the problem?
 - i. What parts did you find helpful?
 - ii. How did they help you solve the problem?

As the usefulness of feedback from bilinguals for evaluating language equivalence has been widely recognized (Streiner & Norman, 1995), French Immersion students were also asked

to compare the two language versions of each item and to look for any nonequivalence between them. After answering the four questions listed above, they were presented with the English version of the item. Upon finishing reading the English version, they were asked:

5. Do the two versions mean exactly the same thing to you? On a 3-point scale—different, similar, identical--how would you rate their comparability in meaning?

If they responded “different” or “similar”, they were then asked:

6. Do you find any differences in wording? If so, where are they and how do the words differ?

If they had found certain parts of the French version confusing earlier, they were asked:

7. Is the confusing part you mentioned in the French version clearer in English? If so, why do you think so?

Structured observation sheet. The structured observation schedule included instructions for the interviewers to record time taken to complete the set of items, and events that were not captured on audiotape, such as the use of gestures.

Interviewers

The researcher conducted the interviews of English-speaking students. Two bilingual interviewers were recruited to interview the French Immersion students. The first interviewer was a French native who immigrated to Canada about 10 years ago. He learned English in high school and took English courses in Canada. With a Bachelor of Education degree, he had taught French as a second language, tourism, and computer studies to junior high and high school students in Canada. The second interviewer was a Canadian Anglophone, who took a French Immersion program from Grade 7 to Grade 12. She had a Bachelor of Arts in French as well as a

Bachelor of Education. She had previously taught elementary and junior high school French Immersion students and also had translation experience between English and French.

The two interviewers were asked to read and sign a confidentiality agreement. The training for the interviewers took two hours, and included a review of the think-aloud procedures, structured interview guidelines, the use of observation sheets, and the use of tape recorders and microphones.

Interview Procedure

Ericsson and Simon (1993) developed a model for verbalization processes of subjects under specific conditions so that inferences could be made about the cognitive processes that produced the verbalization. They made a distinction between two types of verbalization: concurrent and retrospective. Concurrent verbalization involves verbalizing the information one attends to while completing a task. Retrospective verbalization occurs after the task has been completed and involves recollection of one's thought processes. Retrospective reports serve to complement, elaborate, and validate the content of concurrent reports. It is recommended that, whenever appropriate, both concurrent and retrospective reports be collected (Ericsson & Simon, 1993). In this study, both concurrent and retrospective reports were collected.

Each student was trained at the beginning of his/her session with one question taken from the mathematics test if the student was to respond to mathematics items or one question taken from the social studies test if the student was to respond to the social studies items. Each student was asked to talk aloud about what he/she was thinking and what information he/she was attending to while answering each question. Probes in the concurrent portion of the interview were kept to minimum. The students were only to be reminded to keep talking after 5 to 10 seconds of silence. After an answer had been selected, the students were then asked the probe

questions listed above. The students who responded to the English version of the items were expected to report in English and the French Immersion students who responded to the French forms were expected to report in French. However, if a French Immersion student started to think aloud in English, he/she was allowed to do so. The entire process was audio-taped.

Data Analysis

Protocol analysis (Ericsson & Simon, 1993) was used to determine and compare the patterns in which students from the two language groups interpreted the items. The protocols from English-speaking students were transcribed and verified by the researcher. The French Immersion students' protocols were transcribed and translated into English by the female bilingual interviewer. To check the accuracy of her translations, another bilingual person translated and transcribed a sample of the French Immersion students' responses. This person was female, born in England, educated in France, and majored in English at university. She was teaching French at the university level in Canada at the time she participated in the study. Three mathematics and seven social studies student protocols were randomly selected from the corresponding sets of French Immersion interview protocols. The number of social studies protocols selected for this part of the study was greater because social studies had three forms while mathematics had two, meaning that more students were interviewed for social studies. Further, due to the nature of the subject, social studies protocols tended to be longer and more complicated than mathematics. Thus, more protocols were selected for the social studies translation verification. For each of the 10 protocols, two items were randomly selected for the second translation. The researcher compared the two versions of translation, and found a very close fit: the wording might have been different at times, but the meanings stayed the same.

Following verification of the transcripts, the protocols were interpreted and coded for each interview item. The focus of the analysis was to determine for each item: (a) how well the students understood the meaning of the question; (b) what aspects of the question, if any, hindered the students in solving the problem; (c) what aspects of the question, if any, facilitated the students in solving the problem; and (d) to what degree the two language versions were different in meaning or wording. That is, four basic categories, corresponding to each of the themes in the think-aloud protocols were created for each test question: unknown words, understanding of the question, confusing parts, and helpful parts. For the French protocols, one more category was added: comparison of the two language versions of the item.

The coding of the protocols was completed by the researcher (see Appendix C for the coding scheme). Reliability of the coding was then estimated by having two independent raters code a sample of the interview protocols. The two raters were experienced Grade 9 teachers from a French Immersion school. At the time of their participation in the study, one teacher was teaching mathematics and the second teacher was teaching social studies. All of the items used in the think-aloud interviews were first divided into three categories based on their difficulty level of coding: hard, medium, and easy. Then three mathematics and seven social studies items were randomly selected across the three categories. For each selected item, four students were randomly selected for coding out of the eight students interviewed. That is to say, 16.7% of mathematics coded data and 15.9% of social studies coded data were reviewed by the two raters.

The training of the teachers took one and a half hours, including a review of the coding procedure and the coding schemes used in this study. One mathematics and two social studies items were selected for the purpose of training. For each item, eight transcripts (four for each language) were randomly selected. The researcher and the two raters coded the first two

transcripts for each subject together. Then the two raters coded the remaining training protocols independently, followed by discussions of each transcript. Attention was paid to making sure that the teachers had a solid understanding of the coding task they were supposed to accomplish. Following the training session, the two teachers coded the 12 mathematics and 28 social studies transcripts independently.

Inter-coder Agreement

Table 1 contains a summary of the inter-rater agreement. For the English data in both mathematics and social studies, the inter-rater agreement was above 95%. For the French data in mathematics, the agreement between the researcher (Coder 1) and the mathematics teacher (Coder 2) was 100%. The agreement between social studies teacher (Coder 3) and the other two raters was lower (91.7%). This may be attributed to the fact that the third coder specialized in social studies, and did not fully follow students' thought process at times (based on comments by Coder 2). For the French data in social studies, the inter-rater agreements were lower (91.1% to 95.5%), but still satisfactory. This may be due to the interaction of more complex protocols and the addition of a coding theme (comparison of English and French versions). As the two teachers commented, coding of the mathematics data was not at all difficult, but coding of the social studies data was somewhat difficult. The mean inter-rater agreements varied from 93.2% to 97.2%. Taken together, the results revealed that the coding was reliable (Krippendorff, 1980) and valid. Of the items that did not receive 100% agreement, most of the disagreement occurred in coding students' understanding of the question. To resolve any discrepancies that occurred, discussions were held between the researcher and the two teachers until consensus was reached. Since all the discrepancies were question specific, there was no need to modify the coding

schemes, and therefore no changes were made to the coding of other items that were not included in the reliability check.

Results

Mathematics

For the think-aloud interviews in mathematics, nine items (five DIF, four non-DIF) were administered. Eight students in each language group responded to each item. An analysis of student verbal responses suggested that no supporting evidence for DIF was found in four of the five DIF items: Items 8, 9, 11, and 24 (see Table 2). For these items, students from both language groups had little problem understanding the items, and the French Immersion students rated the two language versions the same for all the items. Some evidence was found for the occurrence of DIF in Item 12. What follows next is the analysis results for Item 12 to illustrate the source of DIF, namely, inadequate proficiency in French on the part of French Immersion students.

English

12. Together, three friends have \$256.00. There are three times as many \$5 bills as there are \$20 bills and four fewer \$10 bills than \$5 bills. If there are three times more loonies than \$5 bills, how many \$20 bills are there?

- A. 4
- B. 8
- C. 12
- D. 16

French

12. Ensemble, trois amis ont 256,00 \$. Il y a trois fois plus de billets de 5 \$ que de billets de 20 \$ et 4 billets de 10 \$ de moins que de billets de 5 \$. S'il y a trois fois plus de pièces de 1 \$ que de billets de 5 \$, combien de billets de 20 \$ y a-t-il?

- A. 4
- B. 8
- C. 12
- D. 16

Item 12 displayed B-level DIF favouring English-speaking examinees. This item contained a lot of information for the students to digest. Although no students from either group identified any unknown words, four students from each language group indicated that the question was confusing because it contained too much information. These eight students either answered the question incorrectly or obtained the right answer by guessing. As one English-

speaking student who answered the item incorrectly put it, “there are lots of loops and steps that you have to go through, if you want to find the answer.” Of the eight students who commented about the large amount of information, three English-speaking and one French Immersion students did not completely understand the item. They found it difficult making sense of mathematical relationships such as *three times as many... as* and *four fewer... than....*

Out of the eight French Immersion students who compared the English and French versions, four students considered them identical. The other four students considered them similar, but found the English version clearer. One justification was “because I read more in English and just use French at school.” In particular, two of them thought *four fewer \$10 bills than \$5 bills* was easier and flowed better than the French counterpart *4 billets de 10 \$ de moins que de billets de 5 \$* [four \$10 bills less than \$5 bills]. Upon consultation with the bilingual research team member, the translation of this item was correct, but a high level of reading was required for both language versions. Some French Immersion students found the English version clear because English was their first language and their French experience was mostly restricted within classroom settings. Therefore, the advantage of the English group as suggested in the exhibition of the DIF appears to be attributed to the interaction of load of information and lack of opportunity to read everyday French on the part of French Immersion students.

Social Studies

For the think-aloud interviews in social studies, 22 items (17 DIF, 5 non-DIF) were administered to the students. Eight students responded to each item in either English or French with the exception of items in Form 2 (Items 3, 4, 5, 13, 25, 26, 31, and 39) of the French version. One of the eight students interviewed using Form 2 turned out to be Francophone, so his verbal report was excluded from all subsequent analyses. Further, for the French version of Item

39, only five verbal reports were obtained because the interviewer run out of time. An analysis of student verbal responses revealed some evidence that helped to explain the DIF in eight items (see Table 3).

Supporting Evidence for DIF

Some supporting evidence for DIF was found for 8 of the 17 DIF items: Items 1, 3, 4, 9, 26, 28, 37, and 39. Interestingly, all these items favoured the English group. In terms of possible causes for DIF, these items can be classified into two groups: differential familiarity with key words/concepts (Items 1, 9, 26, 28, and 37), and differential difficulty of stimulus texts (3, 4, and 39).

Differential familiarity with key words/concepts. Due to the limited space in this paper, two items are presented to illustrate the contribution of differential familiarity with key words/concepts to the occurrence of DIF.

Item 9 (as presented below) displayed B-level DIF favouring English-speaking examinees. Four of the eight French Immersion students found the two versions identical. The other four students found the two versions similar in meaning, but the English version was clearer. Three of these four students pointed out in particular that they had trouble making sense of *la chaîne de montage*, while no English-speaking students had difficulty understanding the English counterpart, *assembly line*. Besides, the phrase *ateliers fermés* was found confusing by four French Immersion students, while no English-speaking students had difficulty with the English counterpart *closed shops*. The word *monopolies/monopoles* was also found difficult by two students from each group. Taken together, students' differential familiarity with the key concept *assembly line* as well as the phrase *closed shop* might have led to the exhibition of B-level DIF favouring the English group.

English

9. Which of the following revolutionary practices allowed Henry Ford to produce a good quality car at an affordable price?

- A. Formation of monopolies
- B. Introduction of closed shops
- C. Invention of the assembly line
- D. Cooperation of the trade unions

French

9. Laquelle de ces pratiques révolutionnaires a permis à Henry Ford de créer une voiture de bonne qualité à un prix modéré?

- A. La formation de monopoles
- B. L'introduction d'ateliers fermés
- C. L'invention de la chaîne de montage
- D. La coopération des syndicats

Similarly, Item 37 (as presented below) displayed C-level DIF favouring English-speaking examinees. Six of the eight French Immersion students who were interviewed reported that the two language versions meant the same, although five of the six students found the English text somewhat clearer. The two remaining students considered the two versions similar, but the English version was a lot clearer to them than the French version. No English-speaking students reported problems understanding the key answer *scarcity*, while five French Immersion students had trouble understanding the French counterpart, *la pénurie*. These five French Immersion students all considered the English version somewhat clearer. Two mentioned specifically that they recognized *scarcity* in English but not *la pénurie* in French. Two students switched their answer to A, the correct answer, after reading the English version. That is to say, although the two words mean the same (according to the review by certified translators), *scarcity* seems to be an easier word for the English group than *la pénurie* for the French group. In other words, their differential familiarity with the key answer might have led to the exhibition of C-level DIF favouring the English group.

English

It is not what strangers think of the Perestroïka that is important, say the Soviets, what matters is what happens here. Within the last 5 years, instead of improving, the situation has become worse. The grocery stores offer fewer products and the stores, with their poor quality and their old-fashioned clothes, look more and more like the Salvation Army store.

37. What is the problem identified in this paragraph?

- A. Scarcity
- B. Repression
- C. Corruption
- D. The black market

French

Ce n'est pas ce que les étrangers pensent de la Perestroïka qui est important, disent les Soviétiques, c'est ce qui se passe ici. Depuis 5 ans, loin de s'améliorer, la situation économique du pays s'est aggravée. Les comptoirs d'alimentation offrent de moins en moins de produits et les magasins, avec leur marchandise de mauvaise qualité et leurs vêtements démodés, ressemblent de plus en plus à des comptoirs de l'Armée du Salut.

37. Quel est le problème évoqué dans ce paragraphe?

- A. La pénurie
- B. La répression
- C. La corruption
- D. Le marché noir

To sum up, the above two items shared one thing in common: the French Immersion students were not as familiar with some key words or concepts as their English counterparts. These key words and concepts were crucial, however, in determining the best answer. Therefore, this differential difficulty might have contributed to the occurrence of DIF favouring the English group.

Differential difficulty of stimulus texts. Due to the limited space in this paper, two DIF items along with one non-DIF item are presented next to illustrate the contribution of differential difficulty of stimulus texts to the occurrence of DIF.

As presented below, Items 3 and 4 are two of the three items that were referenced to the same stimulus text. They both displayed C-level DIF favouring English-speaking examinees. The other item (Item 5) did not exhibit DIF. For comparison purpose, Item 5 was also administered in the think-aloud interviews.

English

3. According to this text, we can say that CUPE Local 287
- A. supports the North Battleford city council.
 - B. supports the privatization of the sewer treatment plant.
 - C. opposes the privatization of the sewer treatment plant.
 - D. supports a partnership with public and private sectors in the issue of sewer treatment.
4. According to this text, what is the **most** important issue raised by CUPE Local 287?
- U.S. Filter Canada
- A. will not do a good job.
 - B. is a property of a French multinational.
 - C. is more interested in profit than water quality.
 - D. threatens the job security of the municipal workers.
5. According to this text, CUPE Local 287 is a
- A. group of concerned citizens from North Battleford.
 - B. business specialized in sewage treatment.
 - C. non-profit organization.
 - D. workers' union.

French

3. Selon ce texte, on peut affirmer que la section locale 287 du SCFP
- A. appuie le conseil de ville de North Battleford.
 - B. appuie la privatisation de la station d'épuration des eaux d'égouts.
 - C. s'oppose à la privatisation de la station d'épuration des eaux d'égouts.
 - D. appuie un partenariat public-privé dans le dossier de d'épuration des eaux d'égouts.
4. Selon ce texte, quelle est la préoccupation **principale** de la section locale 287 du SCFP?
- U.S. Filter Canada
- A. ne fera pas un bon travail.
 - B. est la propriété d'une multinationale française.
 - C. est plus intéressée par les profits que par la qualité de l'eau.
 - D. est une menace pour les emplois des employés municipaux.
5. Selon ce texte, la section locale 287 du SCFP est
- A. un groupe de citoyens inquiets de North Battleford.
 - B. une entreprise d'épuration des eaux d'égout.
 - C. une association à but non-lucratif.
 - D. un syndicat de travailleurs.

The text (see Appendix D) used for Items 3, 4 and 5 is one of the longer texts included in the social studies test. The English version is 153 words long, and the French version is 189 words long. In terms of the comparability of the two versions of the text, six out of the seven French Immersion students said that the English version was clearer/easier than the French

version. The reasons underlying their judgments included: “the terms used in French were quite difficult to understand”, “I recognize lots of the terms in English”, “The words [in English] were more familiar to me and I really don’t like reading in French”. While only one of the eight English-speaking students found the text confusing, four of the seven French Immersion students found it hard to understand. When it comes to unknown words, two English-speaking students mentioned the word *municipal*, which they did not think affected how they answered the questions. In contrast, two French Immersion students did not understand *égout* [sewer], which happened to be a more important word for understanding the messages conveyed through this text.

For Item 3, six of the seven French Immersion students considered the meaning of the two versions the same, and one considered them similar. While no English-speaking students had difficulty understanding this item, two French Immersion students reported problems with the key word *appui* [support]. Therefore, students’ differential familiarity with this key word and words in the text in general might have caused their differential performance favouring the English group.

For Item 4, six of the seven students commented on the comparability of the two versions. While four considered them identical, two noted that *the most important issue* and *la préoccupation principale* have similar meanings, but different emphases. To them, *la préoccupation principale* meant *main concern* rather than *the most important issue*. It is not clear to what extent this difference affected how students selected their answer to this question. However, one thing common for Items 3 and 4 is that they both required adequate understanding of pretty much the whole text. For Item 3, the eight English-speaking students and three of the French Immersion students found Paragraph 3, especially the last part, helpful in answering the

question. One French Immersion student acknowledged Paragraph 1 helpful. Similarly, for Item 4, seven English-speaking students and four French Immersion students found Paragraph 3 helpful. Therefore, the DIF for Items 3 and 4 might have been caused by the differential difficulty of the text, because answering these items involves sufficient understanding of the whole text.

In contrast to Items 3 and 4, Item 5 did not display DIF. Six of the seven French Immersion students considered the two versions the same in meaning, while one considered them similar. Analysis of student verbal reports indicated that the word *non-lucratif* [non-profit] in the text was a new word for all the seven students in the French group, while students from the English group had no problem at all understanding this item. In spite of this difference, Item 5 did not produce DIF. For one thing, *non-lucratif* was not in the right option. For another, answering Item 5 required access to local information rather than global information, as compared to Items 3 and 4. All the eight English-speaking students and three of the French Immersion students found one sentence particularly helpful in answering this question: “CUPE Local 287, which represents 123 municipal workers including sewer and water plant operators, outlined their concerns” In other words, this definition of CUPE Local 287 provided key information for answering Item 5.

For the above three items, the stimulus text was more difficult for the French Immersion students than for the English-speaking students. As a result, this differential difficulty might have contributed to the occurrence of DIF favouring the English group on two of the three items.

To summarize, some evidence was found that explained the DIF in 8 of 17 DIF items in social studies. Some French Immersion students’ inadequate proficiency in French might have to

a varying degree contributed to the DIF in eight items. For the other nine items, the DIF might be due to impact (real ability differences), or other unknown factors.

Discussions and Conclusions

The purpose of this portion of the total study was to determine the extent to which the DIF found in the second stage could be attributed to adaptation-related differences. Protocol analysis (Ericsson & Simon, 1993) was used to determine and compare the patterns in which students from the two language groups interpreted the items. Responses of examinees from the think-aloud interviews were used to determine whether the examinees were helped or hindered by certain aspects of the items. Similar to what Ercikan et al. (2004) established, the think-aloud protocol approach was found to be useful in identifying sources of DIF as a complementary tool to approaches like judgmental reviews and statistical methods.

For mathematics, no supporting evidence was identified to have caused the DIF found in four of the five items. The DIF in Item 12 appears to be attributed to the interaction of heavy loads of information and inadequate proficiency in French for French Immersion students. For the 17 DIF items in social studies, no supporting evidence for DIF was found in nine items. For the other eight items, the DIF could be attributed to either differential familiarity with key words/concepts or differential difficulty of stimulus texts, which was caused mostly likely by the inadequate proficiency in French on the part of French Immersion students. The DIF in the other nine items could be attributed to impact or other factors. It is also possible that our think-aloud protocols did not induce the kind of responses from students that would help explain the DIF, or the limitations in the sample we had restricted the kind of responses we would get from them.

As discussed earlier, some key words/concepts in the social studies items were found to be not equally difficult for students from the two language groups. These words/concepts include

mass production (Item 1), assembly line (Item 9), unemployment (Item 26), and scarcity (Item 37). Upon consultation with the bilingual team member of this project, the translation of these key words/concepts was correct. An examination of Alberta Grade 9 social studies provincial tests (1998-2001) revealed that these key words/concepts were part of their tests as well. Their corresponding translations of these key words were the same as those used in this study.

Therefore, it seems reasonable to conclude that the DIF found in these items was not attributed to adaptation-related differences.

Similarly, some stimulus texts were found to be more difficult for the French Immersion students. The text for Items 3, 4 and 5 served as a good example. On one hand, due to the nature of the languages, the French text was longer than the English version. On the other hand, a great majority of the French Immersion students (six out of seven) found the English version clearer and easier. As one student summarized, “the English one [version] was shorter and less complicated. In French they used some vocabulary that you really have to think about.” Therefore, the differential difficulty of the stimulus texts might have contributed to the DIF found in two of the three items referenced to the same text.

Bias or not?

A review of literature on French Immersion education suggested that the differential difficulty of these words/texts could be attributed to French Immersion students’ lack of exposure in French outside the classroom. Day and Shapson (1996) provided a lucid example:

A grade 4 item asking students what they should do first if a piece of bread were caught in a toaster illustrates problems in using translated items for different groups of students. The English version of this item contains words that are familiar to English-speaking children (e.g., toaster, plug, poke); however, the corresponding words in French may not necessarily be known by immersion students because their experiences in French tend to be limited to the classroom. Even though the translation is correct, the items are not equally difficult in the two languages because of the different linguistic experiences of the two groups (p. 16).

Similarly, Romney, Romney, and Braun (1989) found that French Immersion students' knowledge of words related to out-of-school activities was limited and this impeded their reading considerably. Another study of Grade 5 French Immersion students in Alberta (Romney, Romney, & Menziers (1995) indicated that the main difficulty for them to read in French was vocabulary. Besides, Romney, Romney, & Menziers also found that more than two-thirds of the French Immersion students never read at all in French for pleasure outside school. The average amount of time they devoted to reading in French was 25 minutes a week as compared to 183 minutes a week in reading in English. Likewise, they watched considerably less television in French (8 minutes per week) than in English (478 minutes per week).

In addition, French Immersions students' disadvantage in writing French tests as compared to English tests in language intensive areas has been demonstrated in earlier research. As illustrated in the study by Morrison and Pawley (1983), Grade 10 French Immersions students performed less well in history when they were tested in French than when they were tested in English. Similarly, Samuel (1990) documented that French Immersion students were not as able to demonstrate their knowledge and skills in social studies when tested in French as when they were tested in English. When French and English forms of social studies achievement tests were randomly assigned to Grade 6 French Immersion students, those who wrote in French achieved significantly lower scores than those who took the English test. In particular, the effect sizes on topic specific text-based questions were all larger than the effect sizes on the same topic discrete items. Taken together, these studies provided supporting evidence for the contribution of French Immersion students' inadequate proficiency in French to the exhibition of DIF favouring English-speaking students in the items discussed above.

An item is considered biased if it is established that the source of the extra difficulty for one group is not relevant to what the test measures (Camilli and Shepard, 1994). If the bias is caused by adaptation, the item should be revised to make sure that the vocabulary, sentences or text mean the same and are equally difficult for students across the groups. When revision fails to achieve the equivalence across different language versions, the biased item should be deleted. Based on Camilli and Shepard's definition, the items with differential difficulty of words/concepts/texts for English-speaking and French Immersion students should be considered biased because the extra language difficulty for French Immersion students was not relevant to the constructs of Grade 9 mathematics or social studies. Nevertheless, the extra difficulty for French Immersion students appears to be caused by their inadequate proficiency in French rather than inadequate adaptation. First, the test items were written by French Immersion program teachers who were teaching the subject for which they developed items. They were familiar with the curriculum contents as well as the language levels of their students, so that the languages they used approximated the language level of their students. Second, the vocabulary and texts used in the test items were deemed generally comparable by the certified translators. Third, most of the French Immersion students (more than half in most cases) had no problems understanding the items in French, which means that their French proficiency had reached a level they were expected to achieve. Therefore, for the key words/concepts found to be differentially difficult in this study, not much could be done to correct the problem, especially for concepts like mass production and assembly line. For the students who had not attained level they were expected to achieve, attention should instead be paid to increasing their exposure in French both in and out of schools. Emphasis should also be placed on making sure they master the key concepts or words (e.g., assembly line) in French, as one of the coders in this study suggested.

Alternatively, accommodations can be made for the French Immersion students when they take tests in French. First, French Immersion students should be allowed to use dictionary when writing tests in French. Second, footnotes can be added to explain the concepts in plain languages if they are found difficult in the field tests. Third, from the perspective of test administration, French Immersion students should be given the choice of writing the tests in whatever language they prefer. Fourth, both language versions of each item can be provided next to each other on the same page in order to maximum the opportunity for the French Immersion students to understand the items.

Efficacy of Simultaneous Test Development Approach

Simultaneous test development involves the development of tests in more than one language at the same time. Typically, bilingual item writers are recruited for test development in two languages. These items writers should have subject matter knowledge relevant to the assessments at hand as well as understanding of the characteristics of the students for whom the tests are intended for. Based on the results of this study, the simultaneous approach is both efficient and effective in producing equally good tests across the two languages.

Efficiency. The simultaneous approach requires the teachers to write each item in one language, and then immediately translate it into another language. They can then go back and forth to change either language version until they make sure that both versions mean exactly the same. In this study, although most teachers did not have experience in item writing or translation, it did not take long for them to get ready for the task. The training of the teachers took no longer than three hours. These item writers also considered efficiency and speed as one attribute of the simultaneous test development approach.

It took about 20 hours for each item developer to write 30 items in both languages, close to 40 minutes per item on average. The item review and revision for the 90 items took four hours in mathematics (about 10 minutes per item) and nine hours in social studies (about 20 minutes per item). Then item review and revision for the pilot test forms (70 items) took eight hours in each subject (about 20 minutes per item). In addition, five translators each spent about 10 hours reviewing the 58 item in mathematics and 51 items in social studies (about 30 minutes per item). The last round of revision took four hours for each subject, about 10 minutes per item on average. All together, it took about 110 minutes to develop one mathematics item in both languages, and 120 minutes to develop one social studies item in both languages.

With Alberta Education (Guimont, personal communication, November 21, 2005), it typically takes 15 to 20 minutes to write one English item in either mathematics or social studies. Then on average it takes 50 minutes to translate one mathematics item into French, and 80 minutes to translate one social studies item. Then the review and revision of each item takes about 50 minutes to accomplish. Taken together, it takes Alberta Education about 120 minutes to develop one mathematics item in both languages and 150 minutes to develop one social studies item in both languages.

To sum up, there was no indication that using the simultaneous development approach involved longer development time or higher cost than traditional test development methods, which is similar to what Solano-Flores, Trumbull, and Nelson-Barber (2002) found. Although the above calculation of time taken to write one item is a rough estimation, the simultaneous approach appears to be at least as efficient as the traditional methods. By employing bilingual teams of experts to establish the linguistic equivalence of original as well as adapted bilingual tests, the lengthy process of back translation is also bypassed.

Effectiveness. The evidence collected through the item development stage suggested that the simultaneous test development method allowed the influence and integration of information from item writers and reviewers representing different language and cultural groups to affect test development directly. The discussions that took place extended beyond the simple choice of comparable words and phrases to the form of expressions in each language and whether differences in form would be allowed in an attempt to maintain comparable meaning while recognizing the idiomatic differences between the two languages. Both the French and English versions of each test were equally open to modifications. Evidence suggested that item writers were able to give deeper consideration to subtle language and culture issues in the item development process.

The item review conducted by certified translators indicated that a great majority of the items (93.1% in mathematics and 94.1% in social studies) were considered identical or similar in meaning. More DIF was identified in social studies (42.5%) than mathematics (19.2%), which was consistent with what was found with Alberta Grade 9 provincial tests (Gierl et al., 1999). In general, social studies tests involve more vocabulary than mathematics tests, and thus tend to produce more DIF as well. The analysis of the student protocols found no adaptation bias in either mathematics or social studies. French Immersion students' inadequate proficiency in French appears to have contributed to the exhibition of DIF in eight social studies items. Above all, the English and French versions of the mathematics and social studies tests developed using the simultaneous approach were deemed comparable by the translators. Student interview protocols further validated the equivalence of the two language versions in meaning and wording.

The simultaneous approach is easiest to implement when only two languages are involved. When the research involves more than two languages/cultures, the task of recruiting

item writers indigenous to all the languages and cultures may be a difficult challenge. The quality of the end product, however, would justify the effort made to meet the challenge.

Directions for Future Research

Francophone students were not included in the interview sample of this study for practical reasons. For future research, it would be interesting to see how Francophone students interpret the test items in French. It is expected that they would not suffer the same difficulty as the French Immersion students, given that French is their native language. If so, the equivalence of the test items in English and French would be validated. However, if they share some of the difficulties with the French Immersion students on the DIF items, such as lack of familiarity with certain words or concepts, the adaptation part of the test development could be problematic.

Sample size posed another limitation of the current study. In replication of the study, more students, both English- and French- speaking, should be interviewed.

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.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- Day, E. M., and Shapson, S. M. (1996). *Studies in Immersion Education*. Clevedon, UK: Multilingual Matters.
- Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks, CA: Sage.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural research, 1*, 185-216.
- 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.
- 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., Law, D., Arim, R., Domene J., Lacroix, S., & Gagnon, F. (2004). *Identifying sources of DIF using think-aloud protocols: Comparing thought processes of examinees taking tests in English versus in French*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

- Erkut, S., Alarcon, O., Garcia Coll, C., Tropp, L. R., & Vasquez Garcia, H. A. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology, 30*, 206-218.
- Ericsson, K., & Simon, H. (1993). *Protocol analysis: Verbal report data*. Cambridge, MA: MIT Press.
- Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304-312.
- Gierl, M. J. (2000). Construct equivalence of translated achievement tests. *Canadian Journal of Education, 25*, 280-296.
- 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.
- Hambleton, R. K. (1993). Translating achievement tests for use in cross-cultural studies. *European Journal of Psychological Assessment, 9*, 57-68.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into 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.
- Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL:

Scientific Software.

- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hill, CA: Sage.
- Lin, J., & Rogers, W. T. (2005). *Validity of simultaneous approaches to the development of equivalent achievement tests in French and English (Stage II)*. Poster session presented at the Annual Meeting of the National Council on measurement in Education, Montreal, Canada.
- Morrison, F., and Pawley, C. (1983). *Subjects Taught in French*. Tenth Annual Report to the Ministry of Education, Part 1. Ottawa: Ottawa Board of Education Research Centre.
- 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.
- Romney, J., Romney, D., and Braun, C. (1988). The effects of reading aloud in French to immersion children on second language acquisition. *The Canadian Modern Language Review, 45*(3), 530-538.
- Romney, J., Romney, D., and Menzies, H. (1995). Reading for pleasure in French: A study of the reading habits and interests of French immersion children. *The Canadian Modern Language Review, 51*(3), 474-509.
- Samuel, M. J. (1990). *Language of testing effects for academic achievement of French Immersion students*. Unpublished master's thesis, University of Alberta, Canada.

- 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.
- 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.
- Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, 2, 107-129
- Stansfield, C. W., & Kahl, S. R. (1998). *Lessons learned from a tryout of Spanish and English versions of a state assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales* (2nd ed.) Oxford, UK: Oxford University Press.
- 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.
- 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. (1997). Bias and equivalence in cross-cultural assessment: An

Overview. *European Review of Applied Psychology*, 47 (4), 263-279.

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

Table 1

Inter-rater Agreement for the Coding of Interview Data

	Mathematics		Social Studies	
	English	French	English	French
Coder 1 vs. Coder 2	95.8%	100%	97.3%	92.9%
Coder 1 vs. Coder 3	97.9%	91.7%	95.5%	91.1%
Coder 2 vs. Coder 3	97.9%	91.7%	96.4%	95.5%
Mean	97.2%	94.5%	96.4%	93.2%

Table 2

Sources of DIF in Mathematics

Item	DIF Level	Favouring	Adaptation Bias	Inadequate Proficiency in French	Impact /Other
8	C	English			✓
9	C	French			✓
11	C	English			✓
12	B	English		✓	
24	B	French			✓

Table 3

Sources of DIF in Social Studies

Item	DIF Level	Favouring	Adaptation Bias	Inadequate Proficiency in French	Impact /Other
1	B	English		✓	
3	C	English		✓	
4	C	English		✓	
9	B	English		✓	
11	C	French			✓
12	C	French			✓
13	C	French			✓
17	C	French			✓
18	B	French			✓
19	B	English			✓
26	B	English		✓	
28	B	English		✓	
31	B	French			✓
36	C	French			✓
37	C	English		✓	
38	C	English			✓
39	B	English		✓	

Appendix A

Verbal Report Instructions for Social Studies

Hello _____ (the student's name). My name is _____.

How are you today?

Thank you for participating in this study. I am interested in how you understand the social studies questions that appear on a test. To find out about this, I am going to ask you to THINK ALOUD as you work through each question. By think aloud I mean that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you select an answer. It is important that you do not plan out or try to explain to me what you are thinking. Just act as if you are alone in the room speaking to yourself. Please try hard to talk about what you are thinking. If I notice that you have stopped talking, I will remind you to keep talking. Do you understand what I want you to do? *(Please don't talk to the student during Think-Aloud unless he stopped talking for 5 seconds, remind him to "keep talking")*.

Please take your time, and answer the questions as best as you can. Afterwards, I am going to ask you a few questions about your understanding of the item. Before we start, I would like to remind you that I am not going to tell your teacher or your principal how you answered the questions. This study will not affect your mark in social studies. Do you have any questions before we begin?

We'll start with this practice question. *(Feel free to interrupt the subject in this warm-up exercise in order for him to get the idea of THINK ALOUD and how to answer the following questions)*.

Before each question

Please tell me what you are thinking as you answer this question. Please remember to say everything that is going through your mind.

Now there are a few questions that I would like to ask you about the item you just looked at:

1. Are there any words that you don't know in this question, including stimulus material (text in box, if applicable) and the question itself?
2. In your own words, could you tell me what you believe the question is asking? Imagine you are explaining it to a classmate.

- i. Possible follow-up probes: Why do you believe this?
3. Did you find any parts of the question confusing? If so,
 - i. What parts did you find confusing?
 - ii. Why are they confusing?
4. Did any find any parts of the question helpful in solving the problem? If so,
 - i. What parts did you find helpful?
 - ii. How did they help you solve the problem?

Further instruction for French-immersion students:

After a student has finished answering the above questions, you will show him or her the same item in English.

Now I would like you to read the English version of this item, and tell me if it means the same as the item in French. Feel free to mark things on the paper if you need to. Let's start with the stimulus material (if applicable). (*Ask Questions 5 and 6 separately for the stimulus material and each question*)

5. Do the two versions of the item mean exactly the same thing to you? On a scale of different – similar – identical, how would rate their comparability in meaning?
6. (*If "different" or "similar" was chosen*) Do you find any differences in wording? If so, where are they and how do the words differ?
7. (*If the student found some part(s) of the question confusing previously in the French version*) Is the confusing part you mentioned in the French version clearer in English? If so, why do you think so?

At the end of the interview:

_____ (Student's name), thank you for your time and participation in this study.

You have been really helpful to me. Thank you.

Appendix B**Observation Sheet**

Date: _____

Interviewer name: _____

School: _____

Language _____

	Name	Gender	Subject	Time taken for the interview	Non-verbal behaviour*
1					
2					
3					
4					
5					
6					

* Please record any non-verbal behaviours of the student that may be relevant to the purpose of the study.

Appendix C

Coding scheme for Social Studies

Unknown words?	Understanding? (<u>Underline</u> your answer if unknown words in the question were explained to)	Confusing?	Helpful?	Comparison
No. List and <u>underline</u> unknown words from interviewer probing, and what the student think they mean in brackets	T (text): Full Q (question): Full	No	No	<ul style="list-style-type: none"> • S: Same / Identical • AS: Almost the same • SI: similar • KA: Keep answer • CA: Change answer to ...
Yes. List unknown words <ul style="list-style-type: none"> • Write how the students think they mean in brackets 	T: Partial Explain why Q: Partial Explain why	Yes List the confusing part (specific to questions): <ul style="list-style-type: none"> • how the students think they mean in brackets 	<ul style="list-style-type: none"> • OP: Options (list them) • List the helpful part or key words (specific to questions) • SAB: Same as above 	<ul style="list-style-type: none"> • EC: English clearer / easier / flows better • FC: French clearer / easier / flows better • FE: More familiar with English terminology • FF: More familiar with French terminology • EN: English as a native language • FN: French as a native language <p>* More than one of the above may apply, use slash if so.</p>
	T: No/Little Explain why Q: No Explain why			If mentioned, list the specific words that are not equivalent in the two versions.

Appendix D

Use the following information to answer questions 3 to 5.

Municipal workers in North Battleford are urging the city to strongly reject a proposal to enter into a public private partnership to build a new sewage treatment plant.

U.S. Filter Canada, which is owned by the French multinational Vivendi, approached the City of North Battleford with the proposal this spring. The city is currently discussing various ways of financing the construction of a new sewage treatment plant.

CUPE Local 287, which represents 123 municipal workers including sewer and water plant operators, outlined their concerns in a presentation to North Battleford's City Council on June 17:

We feel it is extremely important to provide City Council with information about the dangers of public private partnerships," said local president Barb Plews. "We don't want the community to lose control of such a vital resource like water to a huge multinational corporation that is more interested in reaping profits than in providing good, clean drinking water.

--- adapted and translated from *Syndicat Canadien de la Fonction publique*

1. According to this text, we can say that CUPE Local 287
 - A. supports the North Battleford city council.
 - B. supports the privatization of the sewer treatment plant.
 - C. opposes the privatization of the sewer treatment plant.
 - D. supports a partnership with public and private sectors in the issue of sewer treatment.

2. According to this text, what is the **most** important issue raised by CUPE Local 287?

U.S. Filter Canada

- A. will not do a good job.
- B. is a property of a French multinational.
- C. is more interested in profit than water quality.
- D. threatens the job security of the municipal workers.

3. According to this text, CUPE Local 287 is a
- A. group of concerned citizens from North Battleford.
 - B. business specialized in sewage treatment.
 - C. non-profit organization.
 - D. workers' union.