

Running head: GROUP COMPARABILITY

**Evaluating the Comparability of English- and French-Speaking
Examinees on a Science Achievement Test Administered using
Two-Stage Testing**

Gautam Puhan

Mark J. Gierl

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Paper Presented at the Annual Meeting of the National Council on
Measurement in Education (NCME) at the Symposium entitled, "Test
Adaptations and Translations: Developments and Evaluation Advances"

Chicago, Illinois, U.S.A.

April 22-24, 2003

Evaluating the Comparability of English- and French-Speaking Examinees on a Science Achievement Test Administered using Two-Stage Testing

Achievement tests are administered routinely in multiple languages to students throughout the world. For example, the International Association for the Evaluation of Educational Achievement (IEA) conducted the Third International Mathematics and Science Study in 1995. The tests were administered in 30 different languages to students in 45 participating countries (Hambleton & Patsula, 1998). Similarly, the Organization for Economic Co-operation and Development (OECD) conducted the Programme for International Student Assessment (PISA) in 2000. Tests of reading, mathematical literacy, and scientific literacy were administered in 13 different languages to students in 32 participating countries (Grisay, 2002). Hambleton (1994), Hambleton and Patsula (1998) and Sireci (1997) contend this trend toward multilingual testing will continue because of an increase in the international exchange of tests, a growing demand for credentialing and licensure exams in multiple languages, the cost efficiency in procuring adapted tests compared with constructing new tests, and a growing interest in cross-cultural research.

Increasingly, multilingual tests are also being administered with alternative testing procedures. For example, the School Achievement Indicators Program (SAIP) in Canada uses two-stage testing (TST; Bock & Zimowski, 1998; Lord, 1971;1980; Zimowski, Muraki, Mislevy, & Bock, 1996) to administer achievement tests in science and mathematics to a representative sample of 13-and 16-year-old students from across the country in both of Canada's official languages, English and French. TST is a procedure where tests of varying difficulty in a second-stage are administered to examinees based on their performance from a test in the first-stage. TST has many advantages. For example, TST can be used to evaluate performance using fewer items than conventional tests but with more measurement precision, especially at the ends of the proficiency score scale; TST can yield more reliable test scores in a shorter period of time because examinees receive items that are matched to their ability level; and TST can be used to measure a broad range of ability across diverse groups of examinees in one testing session. These advantages are well documented in paper-and-pencil testing situations with one language group (Bock & Zimowski,

1998; Lord, 1971, 1980). However, the effectiveness of TST in a paper-and-pencil testing situation with multiple language groups has not been studied carefully.

As noted earlier, TST was used by the Council of Ministers of Education in Canada to assess achievement across a diverse group of examinee in two age groups (13- and 16-year olds) and in two languages (English and French) during one test administration for science and mathematics. Scores for the entire sample were then placed on a single proficiency scale so performance could be compared across all age and language groups. *The tests were administered and scored with the implicit assumption that TST produced equivalent results across language groups.*

However, caution must be exercised when using TST in multilingual testing situations because the effectiveness of many translated tests is often questioned (*Standards for Educational and Psychological Testing*, 1999). Angoff and Cook (1988, p. 2) note: "It can hardly be expected without careful and detailed checks, that the translated items will have the same meaning and relative difficulty for the second group as they had for the original group before translation." These problems may be compounded when complex testing procedures, like TST, are used because the two-stage approach may be differentially effective across language groups. Therefore, a thorough analysis of the adaptation and administration process is necessary to ensure that translated items have the same meaning and statistical characteristics across language forms so that the tests yield scores that are comparable and interpretable across language groups (e.g., Allalouf, Hambleton, & Sireci, 1999; Hambleton & Patsula, 1998, 1999; Ercikan, Gierl, McCreith, Puhan, & Koh, 2002; Gierl & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999; Jeanrie & Bertrand, 1999; Reckase & Kuncze, 1999; *Standards for Educational and Psychological Testing*, 1999).

TST involves administering test forms of varying difficulty in the second stage based on the examinees' ability estimate from the first stage. The first-stage test (also called the routing test) is one of the most important features of the TST procedure. If the first-stage test has items with translation differences, then it may not place examinees from different language groups equally well. Therefore, differential item functioning (DIF) analysis should be conducted on the first-stage test to statistically identify and substantively interpret items that might favor one group which, in turn, could lead to misplacement of these examinees in the second-stage test. The second-stage

test must also be monitored for group comparability. If the first-stage test misplaces examinees, then the second-stage test will contain items that are not properly matched to the examinees' ability levels (i.e., the tests will be either too easy or too hard) producing disparate ability distributions between groups. For example, if the first-stage test, administered to two language groups like English- and French-speaking examinees, has items that are biased and favor French-speaking examinees, then more French-speaking examinees will be routed to the test form for a high-ability group even though these examinees should be routed to the test form for a low-ability group. Consequently, the French-speaking examinees will take a more difficult second-stage test and may perform poorly compared to the English-speaking examinees. Therefore, differences between groups must be evaluated on the second-stage test. Because item response theory (IRT) is often used to scale examinees in TST, group comparability can be evaluated using IRT methods by comparing the groups according to their estimated latent ability distributions, test information functions, standard errors of estimate, test characteristic curves, and reliability indices. In short, a thorough evaluation of the comparability between groups for TST requires substantive and statistical analyses of the first- and second-stage tests.

The purpose of the present study is to evaluate the comparability of English and French examinees on a two-stage science achievement test. The SAIP Science achievement test was administered in 1996 and 1999 to a national sample of 13- and 16-year-old students in Canada in both the English and French languages using a two-stage testing procedure. The tests were administered and scored with the implicit assumption that the two language forms were equivalent. Our study is designed to evaluate this assumption. We followed a two-step process. First, we identified and evaluated the sources of differential item functioning (DIF) on the first-stage test when English- and French-speaking examinees were compared. Second, we evaluated the comparability of English- and French-speaking examinees on the second-stage test using IRT methods. This study is relevant for test developers and users in multilingual countries where concerns about the accuracy of the translation process and the comparability of examinees from different language groups should be particularly important, given the popularity of multilingual testing. We present and illustrate the use of methods that help overcome some of the challenges

inherent when comparing examinees who are administered two-stage tests in multilingual settings.

Method

Achievement Test and Student Samples

SAIP achievement tests are designed to measure "what students in Canadian schools are expected to know and be able to do" (Council of Ministers of Education, Canada, 2000, p. 6). The SAIP Science achievement test is designed to assess the knowledge and skills of 13- and 16-year old Canadian students in the following areas: knowledge and concepts of science, nature of science, relationship of science to technology and societal issues, and science inquiry skills. Performance in these content areas is described over five levels of performance with level one indicating very early stages of science literacy, typical of early elementary education, and level five indicating knowledge and skills acquired by a student who has completed a full range of specialized science courses at or near the end of secondary school (Council of Ministers of Education, Canada, 2000, p. 7). The content of the test is determined by assessment and curriculum specialists from universities, content experts, and representatives from non-governmental organizations across Canada. This broad range of knowledge and skills could only be assessed across a diverse groups of examinees with a single test during one administration using an adaptive testing procedure, like TST.

The test consists of a written assessment and a hands-on performance assessment. The written assessment includes both multiple-choice and constructed-response items that measure the acquisition of concepts, procedures, and problem-solving skills. Both multiple-choice and written-response items in the written assessment are scored dichotomously. In the hands-on performance assessment students are required to perform tasks that required them to collect and analyze their own data and apply inquiry skills to practical problems using real materials. Analyses in the current study were only performed using data from the written assessment.

The test was developed in Canada's two official languages, English and French, and the test was designed to be equivalent for both language groups. However, little information is presented in the SAIP technical reports about the bilingual test development process beyond the fact that

both English- and French-speaking test developers were involved in writing the test items, with the intent of eliminating any linguistic bias (Council of Ministers of Education, Canada, 2000, p. 8).

Students writing the science tests in 1996 and 1999 were administered a first-stage test consisting of 12 items at level three. Based on the results of the first-stage test, examinees were assigned to either an easy or difficult second-stage test. Examinees who scored seven or lower on the first-stage test were assigned to the easy second-stage test whereas examinees who scored eight or higher were assigned to the difficult second-stage test. Each second-stage test consisted of 66 items that covered a different combination of achievement levels ranging from one (lowest) to five (highest). The easy second-stage test contained 26 level one items, 26 level two items, and 14 level three items. Conversely, the difficult second-stage test contained 14 level three items, 26 level four items, and 26 level five items. The 14 level three items were the same for the easy and hard second-stage test. The test was scored and the results were reported with the assumption that the first- and second-stage tests works equally well for both English- and French-speaking examinees.

Eight distinct groups of examinees, classified by ability (low and high), age (13- and 16-year olds), and language (English and French), were analyzed. The ability designation was prescribed through the testing procedure (i.e., examinees were categorized as low or high ability depending on their performance on the first-stage test). The low ability group wrote the easy second-stage test and the high ability group wrote the difficult second-stage test. The age designation was prescribed by the groups in the sample. SAIP tests were administered to a national random sample of 13-year old and 16-year old students in Canada. The language designation was prescribed using the first language of the examinees. English-speaking students in the national sample were administered the English form and the French-speaking students were administered the French form.

The SAIP Science achievement test was administered in 1996 and 1999 using the same test items. For the 1996 Science administration, the low ability group was composed of 13-year-old English (n=5 171) and French (n=1 986) examinees and 16-year-old English (n=2 772) and French (n=1 101) examinees. The high ability group was composed of 13-year-old English (n=4

347) and French (n=1 540) examinees and 16-year-old English (n=5 713) and French (n=2 012) examinees. In the 1999 Science administration, the low ability group was composed of 13-year-old English (n=4 431) and French (n=1 549) examinees and 16-year-old English (n=2 178) and French (n=858) examinees. The high ability group was composed of 13-year-old English (n=4 086) and French (n=1 449) examinees and 16-year-old English (n=5 729) and French (n=2 040) examinees. Data from both administrations were analyzed in the current study as a way of cross-validating the findings using the same items across two different samples of students.

Analytical Procedure

Evaluating the effectiveness of a TST procedure for English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 tests requires an analysis of the first and second-stage test.

Comparability of English and French Versions on the First-Stage Test

A comprehensive analysis of the first-stage test was conducted using statistical and substantive methods. Statistical analyses of the first-stage test included using the DIF detection procedure SIBTEST (Shealy & Stout, 1993) to identify items that function differentially for English- and French-speaking examinees. DIF analyses were conducted on each item from the English and French forms of the first-stage test, using the remaining items as the matching subtest.

SIBTEST provides an overall statistical test and a measure of the effect size for each item (\hat{B}_{UNI} is an estimate of the amount of DIF). Roussos and Stout (1996b) adopted the ETS guidelines for identifying DIF (e.g., Zieky, 1993) and applied the results to SIBTEST. They proposed the following guidelines to classify DIF on a single item using SIBTEST: (a) negligible or A-level DIF: Null hypothesis is rejected and the absolute value of $\hat{B}_{UNI} < 0.059$, (b) moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\hat{B}_{UNI}| < 0.088$, and (c) large or C-level DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| \geq 0.088$. These guidelines are used to classify DIF items in the present study using a alpha-level of 0.05 with a non-directional hypothesis test. In all English-French comparisons, items with a B- or C-level rating were considered DIF items. This

decision is based on procedures used in many sensitivity reviews where B- and C-level DIF items are typically identified and scrutinized for potential bias (Zieky, 1993).

For the substantive analyses, test translators (two females and two males who were also junior high school teachers in bilingual schools) were asked to review all items and identified potential translation problems or differences. A translation review process developed by Gierl and Khaliq (2001) was used in the present study. Four bilingual French-English translators completed a blind review of the first-stage test. Three of the four translators were native French speakers and one translator was a native English speaker. The four translators had extensive experience in teaching, ranging from seven to twenty-three years, and all four translators were nominated to participate in this study by the Assistant Director of the Achievement Testing Unit from the Ministry of Education (in the province of Alberta, where this study was conducted). The four translators were asked to evaluate their skills on a 5-point scale ranging from very unconfident (rating of 1) to very confident (rating of 5). Each translator was either confident (rating of 4) or very confident (rating of 5) about his or her knowledge of the curriculum and of shared meanings and cultural specifics between the English and French languages in Canada.

In the review, the four translators first worked separately. They were asked to evaluate the similarities and differences between the English and French test items in the first-stage test, and to identify any translation problems or differences. More specifically, the four translators were asked to specify, for each item, which language group would be favored, identify the reason or reasons for the translation difference, and categorize the reason or reasons using the four sources of the translation errors identified by Gierl and Khaliq (2001). These sources of translation differences included (a) omissions or additions that affect meaning, (b) differences in the words, expressions, or sentence structure of items that are *inherent* to the language and/or culture, (c) differences in the words, expressions, or sentence structure of items that are *not inherent* to the language and/or culture, and (d) differences in item format. The four translators were also asked to create their own categories if they found the sources identified by Gierl and Khaliq (2001) to be insufficient. Once the independent review was completed, the four translators met to discuss their decisions. The meeting allowed each translator to defend his or her decision

for every item and the test translators, as a group, were asked to reach consensus on the items where they disagreed. The review process required two hours and forty-five minutes in order to reach consensus for all items across the four translators.

Comparability of English and French Versions of the Second-Stage Test

All IRT analyses were conducted using BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996). The analysis was carried out in two steps. The first step consisted of estimating the first-stage item parameters and latent distributions. The second step consisted of estimating the link and second-stage item parameters and the latent distributions¹. For the second-stage analysis, the latent distributions estimated in the first-stage analysis were used as the prior distributions for maximum marginal likelihood estimation of the combined first- and second-stage test data. The latent distributions show the proficiencies for each examinee subgroup in this study. These distributions are typically described using the means and standard deviations (SD) of the achievement levels for different groups of examinees.

All BILOG-MG analyses were conducted using the two-parameter logistic (2PL) IRT model. The 2PL IRT model was chosen for two reasons. First, almost 40% of the items used in the second-stage test were constructed-response items where guessing was expected to be minimal. Second, the second-stage test in TST are tailored to the abilities of the examinees. Therefore it was reasonable to assume that the guessing parameter would be low for the multiple-choice items. To evaluate this assumption, the performance of low-scoring examinees (based on their total score) was examined on the most difficult items. The expectation was that the low-scoring examinees would have a low probability of correctly solving the most difficult items if the assumption of no guessing was true. We considered examinees who scored less than one third of the total score in the second-stage tests as low scoring examinees (cf. Ndalichako & Rogers, 1997). The performance for the low scoring examinees for all eight subgroups on the three most difficult items in the second-stage test was evaluated. As shown in Table 1, performance on the

¹ IRT scaling for items in the second-stage test was conducted using the 14 level 3 items common to the easy and difficult second-stage forms. Scaling was necessary so we could establish a common metric for all eight subgroups in our study.

three most difficult items across the eight sub-groups was close to zero suggesting that guessing was minimal on the second-stage tests.

Comparability across language group was evaluated using four IRT procedures. First, the test information function (TIF) was computed from items administered to the English- and French-speaking examinees, and then compared across language group. A TIF difference indicates that items for the English- or French-speaking examinees do not provide the same amount of information which, in turn, might indicate that items on the two language forms are not comparable. Using a 2PL IRT model, the item information function is calculated as

$$I_i(\boldsymbol{q}) = D^2 a_i^2 P_i Q_i,$$

where $D = 1.7$, a_i is the item discrimination parameter, P_i is the probability of an examinee at a given theta (\boldsymbol{q}) level obtaining the correct answer to item i , and $Q_i = 1 - P_i$. The TIF is the sum of the item information functions, given as

$$I(\boldsymbol{q}) = \sum_{i=1}^n I_i(\boldsymbol{q}).$$

Second, the standard error of estimate [$SE(\hat{\boldsymbol{q}})$] was computed from items administered to the English- and French-speaking examinees, and then compared across language group. A $SE(\hat{\boldsymbol{q}})$ difference indicates that items for the English- or French-speaking examinees do not provide the same measurement precision at specific locations on the score scale which might indicate that items on the two language forms are not comparable. Using the 2PL IRT model, the $SE(\hat{\boldsymbol{q}})$ is given as

$$SE(\hat{\boldsymbol{q}}) = \frac{1}{\sqrt{I(\boldsymbol{q})}},$$

where $I(\boldsymbol{q})$ is the test information function for the 2PL IRT model.

Third, the test characteristic curve (TCC) was computed from items administered to the English- and French-speaking examinees, and then compared across language group. A TCC difference indicates that items for the English- or French-speaking examinees do not provide the

same true score estimate which, in turn, might indicate that items on the two language forms are not comparable. Using the 2PL IRT model, the item characteristic curve is calculated as

$$P_i(\mathbf{q}) = \frac{e^{Da_i(\mathbf{q}-b_i)}}{1 + e^{Da_i(\mathbf{q}-b_i)}} ,$$

where $P_i(\mathbf{q})$ is the probability that a randomly chosen examinee with ability \mathbf{q} answers item i correctly, b_i is the difficulty parameter, and a_i is the discrimination parameter. The TCC is the sum of the item characteristic curves,

$$T = \sum_{i=1}^n P_i(\mathbf{q}) .$$

Fourth, the reliability index was computed from items administered to the English- and French-speaking examinees, and then compared across language group. The reliability index is calculated as

$$R_{xx} = 1 - ME ,$$

where R_{xx} is the reliability index and ME is the measurement error of variance. According to Bock and Zimowski (1998), the measurement error of variance for any given test is the reciprocal of the average test information function estimated across the entire theta scale. Reliability indices between the range of 0.80 to 0.90 are considered appropriate for low-stakes examinations (Bock & Zimowski, 1998, p. 41). A R_{xx} difference indicates that items for the English- or French-speaking examinees are not equally reliable which might indicate that items on the two language forms are not comparable.

The TIF, $SE(\hat{\mathbf{q}})$, and TCC for the second-stage tests were compared across language groups using graphical procedures and statistical tests. All graphical analyses were conducted by comparing the functions between language groups. This approach, however, provided only an estimate of the magnitude of the difference between the groups. To evaluate the graphical analyses, statistical tests were computed to obtain a more precise measure of the difference

between the groups. The mean square residual (MSR) was calculated for each TIF, $SE(\hat{\mathbf{q}})$, and TCC difference using the equation

$$MSR = \frac{\sum_{i=1}^n [X_i(\mathbf{q}) - Y_i(\mathbf{q})]^2}{n-1},$$

where n represents the number of quadrature points on the theta scale, $X_i(\mathbf{q})$ corresponds to \mathbf{q} for the English-speaking examinees, and $Y_i(\mathbf{q})$ corresponds to the \mathbf{q} for French-speaking examinees. The null and alternative hypotheses are

$$H_0 : MSR = 0 \text{ versus } H_1 : MSR \neq 0.$$

The value of the MSR was compared to the critical value in a chi-square distribution with $n-1$ degrees of freedom to test whether the MSR is statistically different from 0 using an alpha level of 0.05.

Results

Two sets of analyses were conducted. First, items on the first-stage test were evaluated for DIF across English- and French-speaking examinees. Second, performance differences on the second-stage tests were compared for English- and French-speaking examinees using IRT-based procedures, as described earlier.

Analysis of the First-Stage Test

DIF analyses of the first-stage test were conducted to identify items that functioned differentially between English- and French-speaking examinees. Initially, we were concerned that the matching subtest may become contaminated if large numbers of DIF items were found in the first-stage test. To overcome this potential problem, we increased the number of items in the matching subtest by adding the 14 items that were common to the easy and difficult second-stage tests to our DIF analysis. Since these 14 items were written by all examinees, inclusion of these items increased the total test score for the matching subtest from 12 to 26, leading to a more reliable matching subtest for the DIF analyses. The DIF items identified using the 12-item first-stage test and the 26-item composite were identical. Hence, the DIF items identified using the 26-item matching subtest are reported. For the 1996 administration, three of the 12 items (items 1, 2,

and 6) were identified as DIF items for English- and French-speaking examinees. Of these three items, two items favored French-speaking examinees and one item favored English-speaking examinees. For the 1999 administration, the same three items were identified as DIF items for English- and French-speaking examinees. The results are shown in Table 2.

When the items were scrutinized during the substantive review, the four translators failed to identify any translation errors in the three DIF items. Instead, these items were judged to be equivalent for examinees in both language groups. The translators did, however, identify two items as potentially problematic, items 3 and 8. Item 3 was believed to favor English-speaking examinees due to differences in words and expressions inherent to the language or culture. Item 8 was also believed to favor English-speaking examinees due to differences in words and expressions not inherent to the language or culture. However, these two items were not statistically significant in the SIBTEST DIF analysis. Consequently, we conclude that items on the first-stage test were free from flagrant translation problems or differences.

Analysis of the Second-Stage Test

Estimated Latent Distributions

For the 1996 administration, the estimated latent distributions for 13-year old, English- and French-speaking examinees within the low- and high-ability groups and 16-year old, English- and French-speaking examinees within the low- and high-ability groups are presented in Figure 1 (Panels A and B, respectively). The means and standard deviations are presented in Table 3. The latent distributions for the low-ability examinees are shifted to the left of the proficiency scale whereas the latent distributions for the high-ability examinees are shifted to the right of the proficiency scale. When low and high ability distributions are compared, there is considerable overlap indicating that the first-stage test did not provide strong separation between the two ability levels for either language group (see Bock & Zimowski, 1998, p. 42 for a contrasting example). The latent distributions between language group but *within ability level*, by comparison, were quite comparable indicating that the first-stage test was routing English and French examinees in a similar, albeit somewhat ineffective, manner. Similar results were found in the 1999 administration when the estimated latent distributions for 13-year old, English- and French-

speaking examinees within the low- and high-ability groups and 16-year old, English- and French-speaking examinees within the low- and high-ability groups were compared (see Figure 1, Panels C and D, respectively and Table 3).

The inadequacy of the first-stage test to separate examinees may be attributed, in part, to the items on the test. The a- and b-parameter estimates for the 1996 and 1999 administration are presented in Table 4. For 1996, the b-parameter estimates ranged from -2.537 to 1.775 ($\bar{X}_B = -0.385$, $SD_B = 1.097$) indicating the items measured a wide range of ability but, as shown with the a-parameter estimates, these items only provided moderate discrimination ($\bar{X}_A = 0.568$, $SD_A = 0.155$) within this range. Hence, some misplacement would likely occur. Similar results were found with the 1999 administration: The b-parameter estimates ranged from -2.290 to 1.594 ($\bar{X}_B = -0.480$, $SD_B = 1.016$) indicating the items measured a wide range of ability but only provided moderate discrimination ($\bar{X}_A = 0.580$, $SD_A = 0.150$) within this range.

Test Information Functions for English and French Versions of the Tests

For the 1996 administration, the test information functions (TIFs) for the English and French versions of the test for 13-year-old, low- and high-ability examinees and 16-year-old, low- and high-ability examinees are presented in Figure 2 (Panels A and B, respectively). Bock and Zimowski (1998, p. 19) recommend that information values range between 5 and 10 for the proficiency score scale spanned by each second-stage test (i.e., the second-stage test associated with the low- and high-ability examinees in the current study). Using these interpretative guidelines, the second-stage test for the low-ability 13-year-old, and 16-year-old examinees across language groups was acceptable along most of the proficiency score scale (i.e., between $q = -3$ to $q = 2$). Conversely, the second-stage test for the high-ability 13-year-old, and 16-year-old examinees was acceptable at the high but not at the low ability range (i.e., below $q = -1.5$).

When the TIFs for the English and French versions of the test were compared *within ability group*, differences appeared for both 13- and 16-year-old, low and high ability examinees. These results might indicate that items for the English- and French-speaking examinees were not

comparable. However, as shown in Table 5, the MSR values for all four comparisons were not statistically significant ($p > 0.05$) indicating that the TIFs were statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

For the 1999 administration, the TIFs for the English and French versions of the test for 13-year-old, low- and high-ability examinees and 16-year-old, low- and high-ability examinees are presented in Figure 2 (Panels C and D, respectively). Using the Bock and Zimowski (1998) guidelines, the second-stage test for the low-ability 13-year-old, and 16-year-old examinees across language groups was acceptable at the low end of the proficiency score scale but not the high end (i.e. above $q = 1.5$). Conversely, the second-stage test for the high-ability 13-year-old, and 16-year-old examinees was not acceptable at the low end (i.e. below $q = -1.5$) of the proficiency score scale but it was acceptable at the high end.

When the TIFs for the English and French versions of the test were compared within ability group, differences, again, appeared for both 13- and 16-year-old, low and high ability examinees. As shown in Table 5, however, the MSR values for all four comparisons were not statistically significant ($p > 0.05$) indicating that the TIFs were statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

SE(\hat{q}) for English and French versions of the Tests

For the 1996 administration, the standard error of estimate [$SE(q)$] for the English and French versions of the test for 13-year-old, low- and high-ability examinees and 16-year-old, low- and high-ability examinees are presented in Figure 3 (Panels A and B, respectively). While the TIF provides a measure of how much information a test provides, the $SE(q)$ provides a summary of the measurement precision for a test. Given that Bock and Zimowski (1998, p. 19) recommended that information values between 5 and 10 were desirable, the comparable standard error ranges from, approximately, 0.32 to 0.45. Using this range as a guide, the second-stage test for the low-ability 13-year-old, and 16-year-old examinees across language groups was acceptable along the majority of the proficiency score scale (i. e., $q = -3$ to $q = 1.5$). Conversely, the second-stage

test for the high-ability 13-year-old, and 16-year-old examinees was acceptable at the high but not at the low ability range on the proficiency score scale (i. e., below $q = -1$).

When the $SE(q)$ for the English and French versions of the test were compared within ability group, differences appear for both 13- and 16-year-old, low and high ability examinees. These results might indicate that items for the English- and French-speaking examinees were not comparable. But, as shown in Table 5, the MSR values for all four comparisons were not statistically significant ($p > 0.05$) indicating that the $SE(q)$ was statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

For the 1999 administration, the $SE(q)$ for the English and French versions of the test for 13-year-old, low- and high-ability examinees and 16-year-old, low- and high-ability examinees are presented in Figure 3 (Panels C and D, respectively). Using the guidelines adapted from Bock and Zimowski (1998), the second-stage test for the low-ability 13-year-old, and 16-year-old examinees across language groups was acceptable at the low end of the proficiency score scale but not the high end (i. e., above $q = 1$). Conversely, the second-stage test for the high-ability 13-year-old, and 16-year-old examinees was not acceptable at the low end of the proficiency score scale (i. e., below $q = -2$) but it was acceptable at the high end.

When the $SE(q)$ for the English and French versions of the test were compared within ability group, differences appear for both 13- and 16-year-old, low and high ability examinees. But as the results in Table 5 reveal, the MSR values for all four comparisons were not statistically significant ($p > 0.05$) indicating that the $SE(q)$ were statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

From our analyses of the TIFs and the $SE(q)$, we conclude that the second-stage tests are not equally effective for the low- and high-ability examinees. The items administered to the low-ability examinees provide more information and measurement precision across a larger range of the proficiency score scale. The items administered to the high-ability examinees, by comparison, provide less information and lower measurement precision across a smaller range of science proficiency. Despite these differences between the ability groups, the differences for

English- and French-speaking examinees within ability group appear small and insignificant as the items provide the same amount of information and measurement precision for examinees in both language groups.

TCCs for English and French Versions of the Tests

For the 1996 administration, the TCCs for the English and French versions of the test for 13-year-old, low and high ability examinees and for 16-year-old, low and high ability examinees are presented in Figure 4 (Panels A, B, C, and D, respectively). The TCCs for the English and French versions of the test appears to be slightly different for both low- and high-ability ability, 13-and 16-year-old examinees. This discrepancy could result in different true score estimates when the language groups are compared, and it might indicate that items for the English- and French-speaking examinees were not comparable. However, as shown in Table 5, the MSR value are similar across the four groups indicating that the TCCs were not statistically different for the English- and French-speaking examinees.

Comparable results were found in the 1999 administration when the TCCs for 13-year old, English- and French-speaking examinees within the low- and high-ability groups and 16-year old, English- and French-speaking examinees within the low- and high-ability groups were compared (see Figure 5, Panels A, B, C, and D, respectively and Table 5).

Reliability Indices for English and French versions of the Tests

The reliability indices are presented in Table 6 for the 1996 and 1999 administrations. The reliability indices for the English and French versions of the test were high (ranging from 0.870 to 0.919) for all eight subgroups, relative to the values suggested by Bock and Zimowski (1998, p.41). Moreover, the R_{xx} outcomes for the English and French versions of the test were comparable within ability group, indicating that the items were equally reliable for English- and French-examinees across age in both the 1996 and 1999 test administrations.

Conclusions and Discussion

The purpose of the present study was to evaluate the comparability of English- and French-speaking examinees on a two-stage science achievement test. The SAIP Science achievement test was administered in 1996 and 1999 to a national sample of 13-and 16-year-old students in

Canada in both the English and French languages using a two-stage testing procedure. The tests were administered and scored with the implicit assumption that the two language forms were equivalent. Our study was designed to evaluate this assumption. A two-step process was used where, first, we identified and evaluated the sources of differential item functioning (DIF) on the first-stage test when English- and French-speaking examinees were compared and, second, we evaluated the comparability of English- and French-speaking examinees on the second-stage test using IRT methods.

Our analyses of the first-stage test indicated that five of twelve items could have translation errors: Three items were identified statistically using SIBTEST and two items were identified substantively by the test translators. When the test review was conducted, the three DIF items were judged to be equivalent by all four translators whereas the two items identified by the translators as potentially problematic showed only small amounts of DIF using SIBTEST. We concluded that the first-stage items were free from flagrant translation problems. The uninterpretable DIF results may be accounted for by two different factors. First, the DIF items, in fact, contain no translation errors. Rather, the differences in performance identified in the statistical analysis reflect legitimate differences between English and French examinees (i.e., item impact). Since impact is not considered a negative attribute, these DIF items may be considered a valid measure of science proficiency on the first-stage test. Second, the translators may have failed to identify sources of group differences because of the instructions used in the test review. Recall, the translators were asked to focus on potential translation problems or difference in the items, not actual group differences between the English- and French-speaking examinees. Therefore, further studies are needed to identify the sources of performance difference--attributable to both bias and impact--on the first-stage test.

Our analyses of the second-stage test indicated that the first- and second-stage tests worked well for some age and ability groups but not others. Despite this outcome, the results were quite similar between language groups. For example, when the latent distributions for the low and high ability distributions were compared, there was considerable overlap between the distributions indicating that the first-stage test did not provide strong separation between the two ability levels

for either language group in either the 1996 or 1999 administrations. This outcome may be attributed, in part, to the moderate a-parameter estimates for the first-stage items. Although the b-parameter estimates reveal that the items measured a wide range of abilities, the a-parameter estimates were only moderate for the 1996 and 1999 administrations indicating that the items did not provide strong discrimination for all examinees in this ability range. Hence some misplacement could occur. Because the first-stage test has such an important effect on the second-stage results, items on the first-stage test must be highly discriminating (e.g., a-parameter estimates of, approximately, 1.0; see Bock & Zimowski, 1998, p. 31) across a range of ability. On the other hand, the latent distributions between language group but *within ability level* were quite comparable for both the 1996 and 1999 administrations indicating that the first-stage test was routing English and French examinees in a similar manner (see Figure 1).

When the TIFs and the $SE(q)$ for the 1996 administration were compared using the interpretative guidelines presented by Bock and Zimowski (1998), the second-stage test for the low-ability 13-year-old, and 16-year-old examinees across language groups was acceptable along most of the proficiency score scale but the second-stage test for the high-ability 13-year-old, and 16-year-old examinees was acceptable at the high but not at the low ability range. Similarly, when the TIFs and the $SE(q)$ for the 1999 administration were compared, the second-stage test for the low-ability 13-year-old, and 16-year-old examinees across language groups was acceptable at the low end of the proficiency score scale but not the high end whereas the second-stage test for the high-ability 13-year-old, and 16-year-old examinees was not acceptable at the low end of the proficiency score scale but it was acceptable at the high end.

Yet when the TIFs and the $SE(q)$ for the English and French versions of the test were compared *within ability group*, the MSR values for all comparisons were not statistically significant ($p > 0.05$) indicating that these differences were negligible for the English and French versions of the tests for 13- and 16-year-old, low or high ability examinees.

When the TCCs for the English and French versions of the test were compared for the 1996 and 1999 administrations, small differences were apparent for low- and high-ability ability, 13- and 16-year-old examinees. However, MSR values were similar across all eight four groups

indicating that the TCCs were not statistically different for the English- and French-speaking examinees. When the reliability indices for the English and French versions of the test were compared for the 1996 and 1999 administrations, the indices were high for all eight subgroups and comparable within ability group indicating that the items were equally reliable for English- and French-examinees across age in both test administrations.

These results allow us to conclude that the second-stage test was not equally effective for the low- and high-ability examinees. The items administered to the low-ability examinees provide more information and measurement precision across a larger range of the proficiency score scale while the items administered to the high-ability examinees provide less information and lower measurement precision across a smaller range of science proficiency. Despite these differences between the ability groups, the differences for English- and French-speaking examinees within ability group appear small and insignificant, and the items provide the same amount of information and measurement precision for examinees in both language groups. Moreover, items on the two language form yield comparable true score estimates for English- and French-speaking examinees (see Figures 4 and 5) and comparable reliability estimates.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement*, 36, 185-198.
- Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the prueba de aptitud academica and the scholastic aptitude test* (College Board Report No. 88-2). New York: College Entrance Examination Board.
- Bock, R. D., & Zimowski, M. F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP*. American Institutes for Research, CA.
- Council of Ministers of Education, Canada (2000). *Report on science assessment, School achievement indicators program, 1999*. Toronto, ON: Council of Ministers of Education, Canada.
- Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2002, May). *Comparability of English and French Versions of SAIP for reading, mathematics and science Items*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Toronto.
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement*, 38, 164-187.
- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.
- Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. In Ray Adams and Margaret Wu (Eds.), *PISA 2000 Technical Report* (pp. 57-70). Paris: Organization for Economic Co-operation and Development.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-224.

- Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45*, 153-171.
- Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*, 1-30.
- 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.
- Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227-242.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Ndalichako, J. L. & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement, 57* (4), 580-589.
- Reckase, M. D., & Kuncze, C. (1999, April). *Translation accuracy of a technical credentialing examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.
- 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.
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16*, 12-19.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BLOG-MG: Multiple-group IRT analysis and test maintenance for binary Items*. Chicago: Scientific Software International.

Author Notes

Gautam Puhan, Associate Measurement Statistician, Educational Testing Service,
Rosedale Road, MS 17-L, Princeton, NJ, 08541

Mark J. Gierl, Associate Professor, Centre for Research in Applied Measurement and
Evaluation, Department of Educational Psychology, 6-110 Education North, Faculty of
Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5

Please address all correspondence to Gautam Puhan, Educational Testing Service, Rosedale
Road, MS 17-L, Princeton, NJ, 08541.

Table 1.

Results of Item-Guessing Analysis for Eight Subgroups on the Second-Stage Test

	1996				1999			
	13-Year Old		16-Year Old		13-Year Old		16-Year Old	
	EN	FR	EN	FR	EN	FR	EN	FR
Low Ability	0.003	0.009	0.002	0.000	0.031	0.059	0.123	0.083
High Ability	0.057	0.004	0.006	0.003	0.014	0.010	0.014	0.087

Note. Each cell represents the average probability of a correct response for the low ability examinees on the three most difficult items on the second-stage test.

EN=English-Speaking Examinees, FR=French-Speaking Examinees

Table 2.

Results of SIBTEST DIF Analysis using the First-Stage Test Items

Item	1996				1999			
	13-Year Old		16-Year Old		13-Year Old		16-Year Old	
	\hat{b}_{UNI}	Level	\hat{b}_{UNI}	Level	\hat{b}_{UNI}	Level	\hat{b}_{UNI}	Level
1	-0.144*	C	-0.087*	B	-0.154*	C	-0.072*	B
2	-0.217*	C	-0.170*	C	-0.225*	C	-0.141*	C
6	0.171*	C	0.147*	C	0.131*	C	0.111*	C

Note. A negative \hat{b}_{UNI} favors the French-speaking examinees.

* $p < 0.05$

Table 3.

Means and Standard Deviations for the Latent Distributions for the Eight Subgroups in the Second-Stage Test

		1996				1999			
		13-Year Old		16-Year Old		13-Year Old		16-Year Old	
		EN	FR	EN	FR	EN	FR	EN	FR
Low Ability	Mean	-0.77	-0.81	-0.52	-0.51	-0.89	-0.97	-0.60	-0.59
	SD	0.78	0.73	0.75	0.71	0.74	0.66	0.71	0.67
High Ability	Mean	0.38	0.26	0.74	0.65	0.34	0.15	0.77	0.62
	SD	0.76	0.74	0.78	0.80	0.70	0.68	0.75	0.76

EN=English-Speaking Examinees, FR=French-Speaking Examinees

Table 4.

Means and Standard Deviations for the IRT a- and b-parameters Estimates in the First-Stage Test

Item	1996		1999	
	a	b	a	b
1	0.366	-0.821	0.372	-0.978
2	0.681	0.089	0.795	0.163
3	0.664	-0.359	0.691	-0.341
4	0.452	-2.537	0.503	-2.290
5	0.253	1.775	0.262	1.594
6	0.612	-0.498	0.672	-0.882
7	0.786	-0.885	0.697	-1.100
8	0.490	-0.364	0.517	-0.640
9	0.582	-0.773	0.573	-0.797
10	0.576	-1.255	0.584	-1.197
11	0.741	-0.111	0.689	-0.159
12	0.615	1.125	0.605	0.862
Mean	0.568	-0.385	0.580	-0.480
SD	0.155	1.097	0.150	1.016

Table 5.

χ^2 -Test Results for the Mean Square Residuals Across the English and French TIFs, TCCs, and SE(θ) Comparisons

	1996						1999					
	13-Year Old			16-Year Old			13-Year Old			16-Year Old		
	TIF	TCC	SE(θ)	TIF	TCC	SE(θ)	TIF	TCC	SE(θ)	TIF	TCC	SE(θ)
Low Ability	1.59	0.30	0.00	1.24	0.90	0.00	0.07	0.11	0.00	0.26	0.02	0.00
High Ability	0.77	0.58	0.00	1.30	0.98	0.00	2.27	1.67	0.00	3.23	0.90	0.00

Note. The critical value was $\chi^2_{\alpha=0.05, 30} = 14.95$.

Table 6.

Reliability Indices for English and French Versions of the SAIP Second-Stage-Tests

	1996				1999			
	13-Year Old		16-Year Old		13-Year Old		16-Year Old	
	EN	FR	EN	FR	EN	FR	EN	FR
Low Ability	0.899	0.901	0.900	0.911	0.913	0.913	0.911	0.919
High Ability	0.870	0.885	0.877	0.893	0.883	0.883	0.888	0.889

EN=English-Speaking Examinees, FR=French-Speaking Examinees

Figure Caption

Figure 1.

Panel A. Latent distributions for 13-year-old, low- and high-ability, English- and French-speaking examinees: 1996 administration.

Panel B. Latent distributions for 16-year-old, low- and high-ability, English- and French-speaking examinees: 1996 administration.

Panel C. Latent distributions for 13-year-old, low- and high-ability, English- and French-speaking examinees: 1999 administration.

Panel D. Latent distributions for 16-year-old, low- and high-ability, English- and French-speaking examinees: 1999 administration.

Figure 2.

Panel A. TIFs for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1996 administration.

Panel B. TIFs for the English and French versions of the test for 16-year-old, low and high-ability examinees: 1996 administration.

Panel C. TIFs for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1999 administration.

Panel D. TIFs for the English and French versions of the test for 16-year-old, low and high-ability examinees: 1999 administration.

Figure 3.

Panel A. $SE(\hat{q})$ for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1996 administration.

Panel B. $SE(\hat{q})$ for the English and French versions of the test for 16-year-old, high and low-ability examinees: 1996 administration.

Panel C. $SE(\hat{q})$ for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1999 administration.

Panel D. $SE(\hat{q})$ for the English and French versions of the test for 16-year-old, high and low-ability examinees: 1999 administration.

Figure 4.

Panel A. TCCs for the English and French versions of the test for 13-year-old, low-ability examinees: 1996 administration.

Panel B. TCCs for the English and French versions of the test for 13-year-old, high-ability examinees: 1996 administration.

Panel C. TCCs for the English and French versions of the test for 16-year-old, low-ability examinees: 1999 administration.

Panel D. TCCs for the English and French versions of the test for 16-year-old, high-ability examinees: 1999 administration.

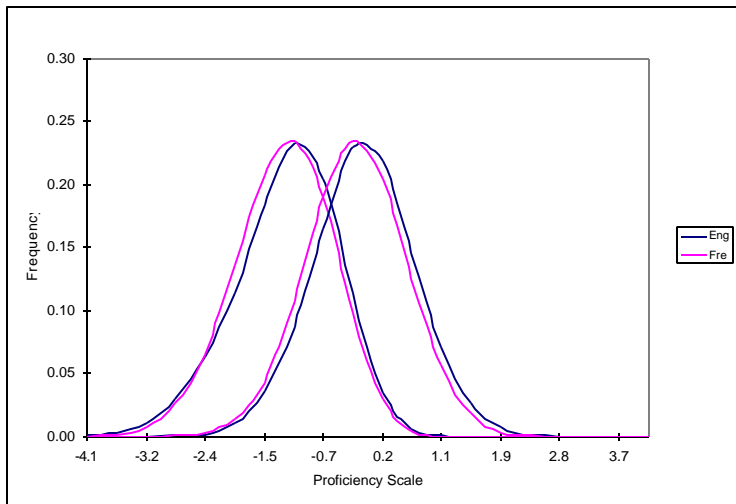
Figure 5.

Panel A. TCCs for the English and French versions of the test for 13-year-old, low-ability examinees: 1999 administration.

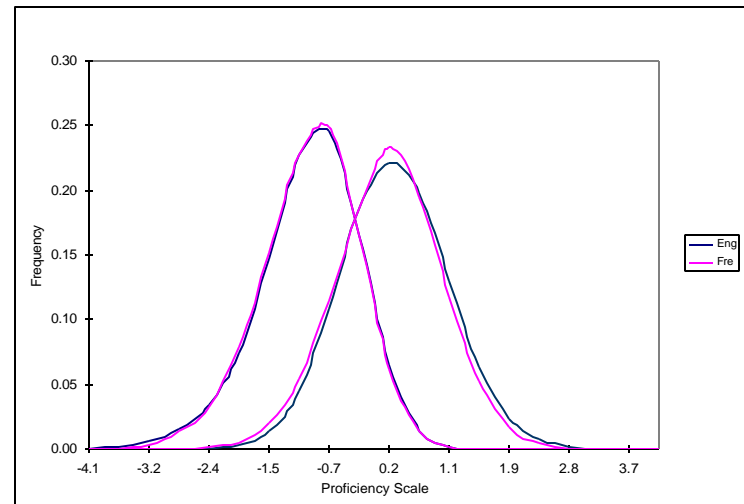
Panel B. TCCs for the English and French versions of the test for 13-year-old, high-ability examinees: 1999 administration.

Panel C. TCCs for the English and French versions of the test for 16-year-old, low-ability examinees: 1999 administration.

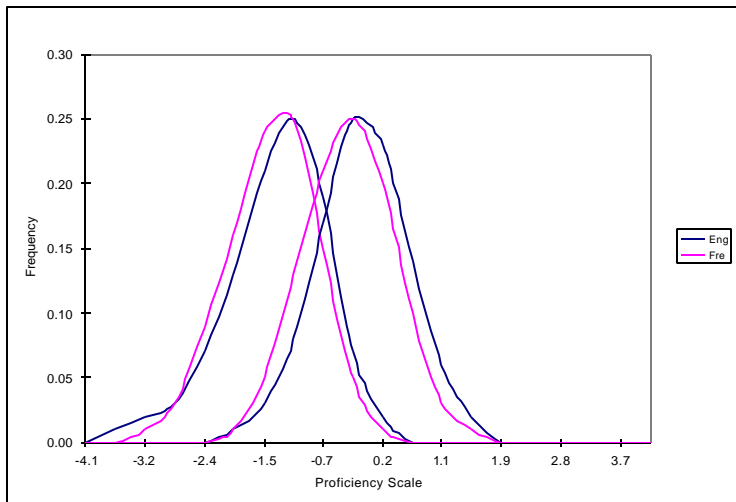
Panel D. TCCs for the English and French versions of the test for 16-year-old, high-ability examinees: 1999 administration.



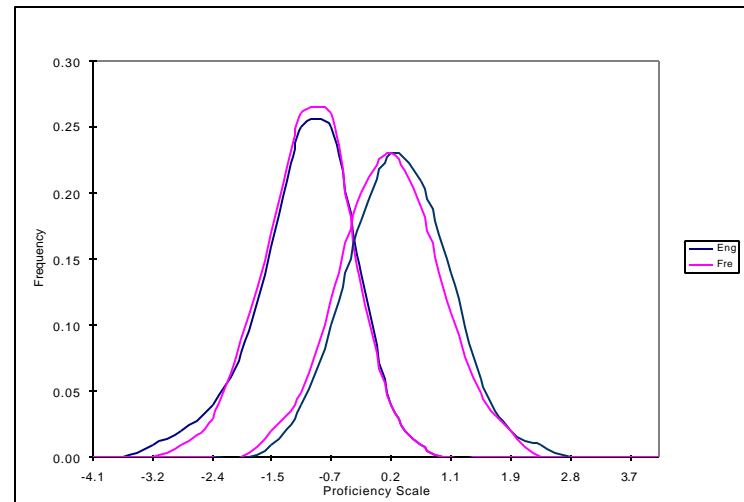
Panel A



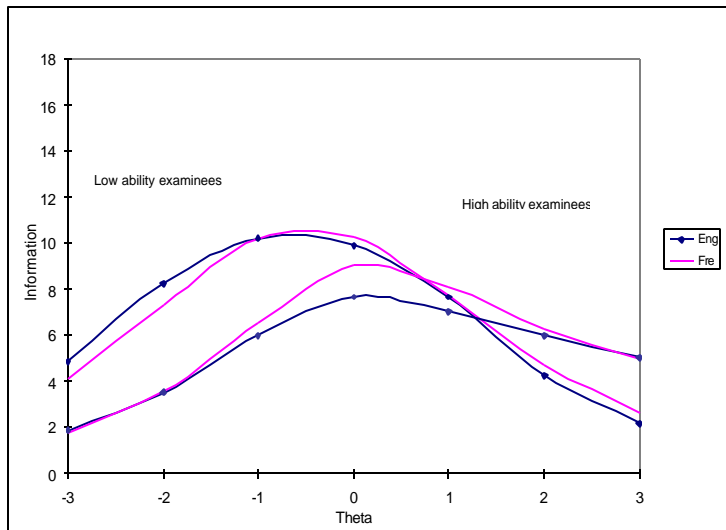
Panel B



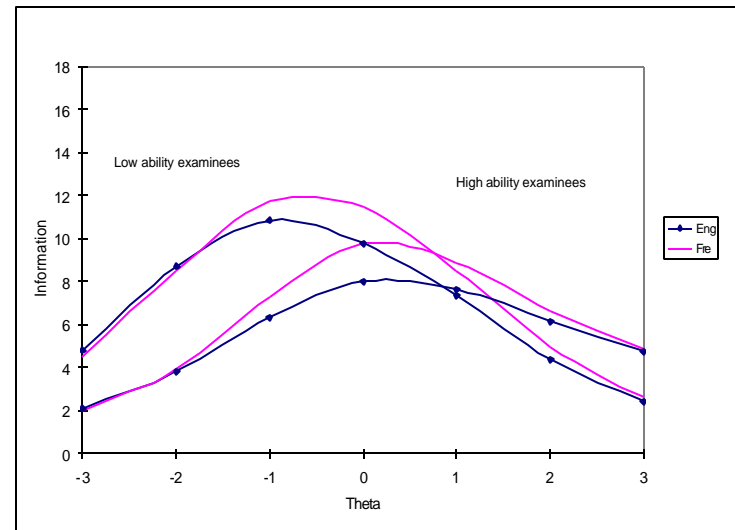
Panel C



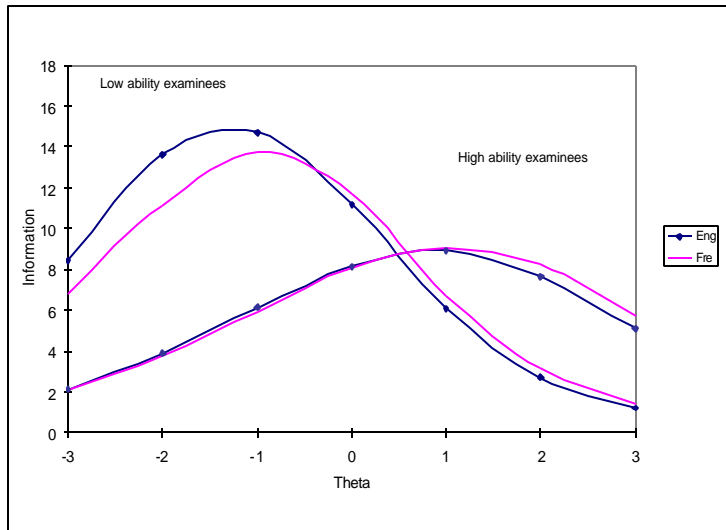
Panel D



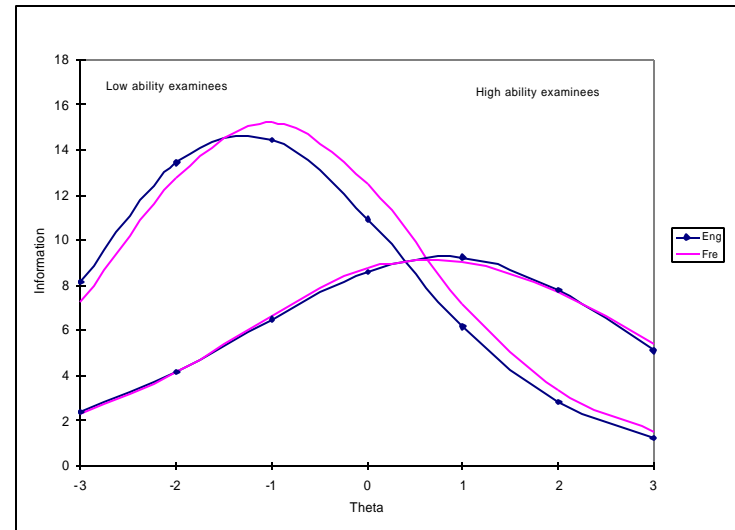
Panel A



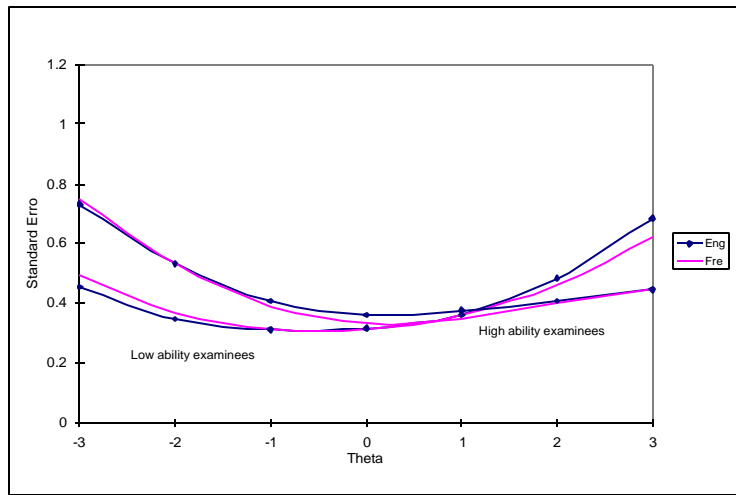
Panel B



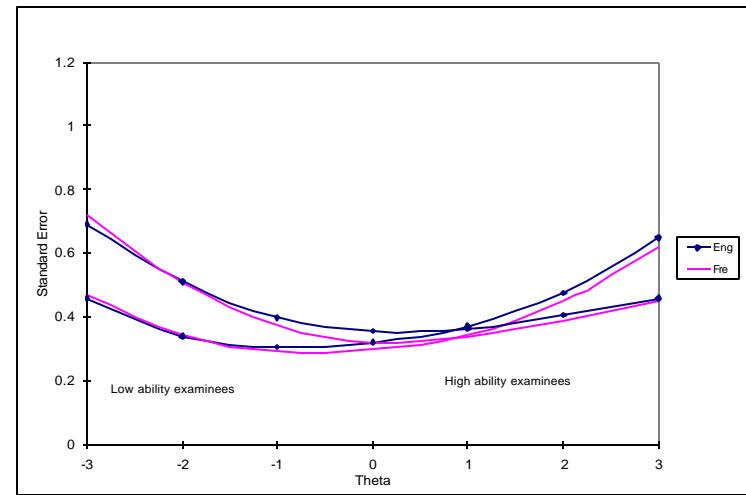
Panel C



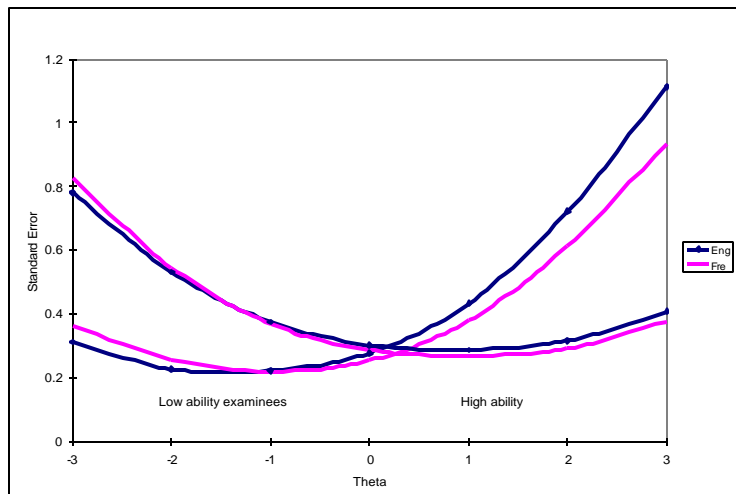
Panel D



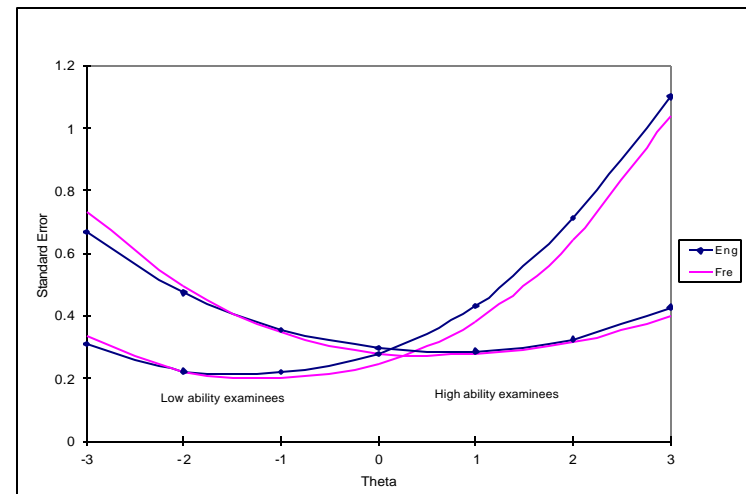
Panel A



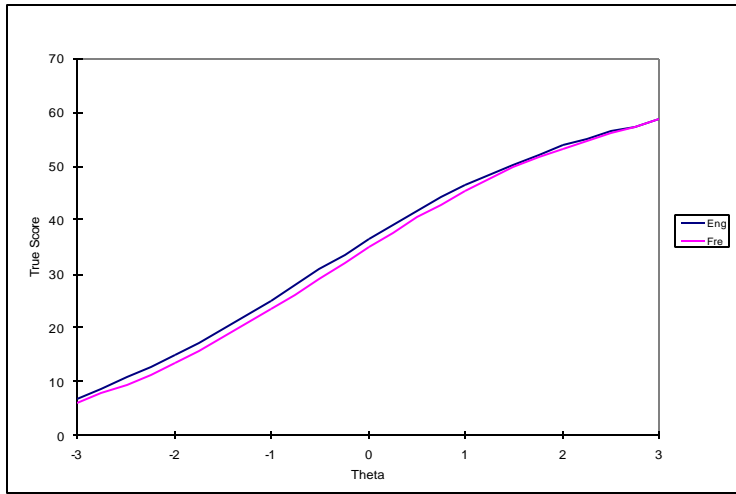
Panel B



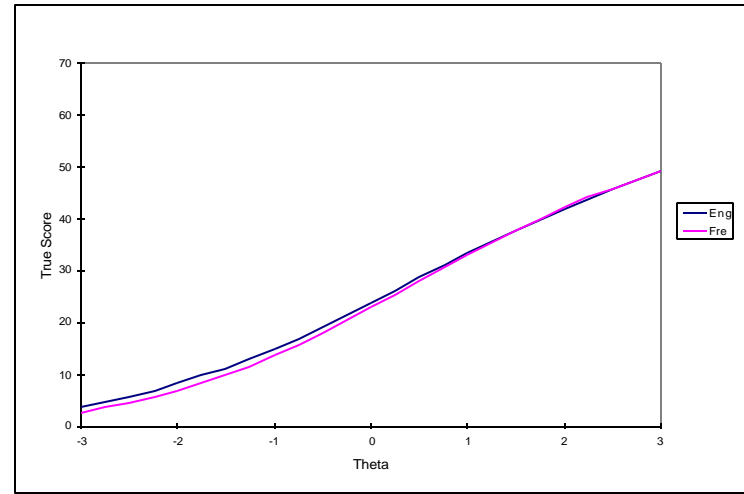
Panel C



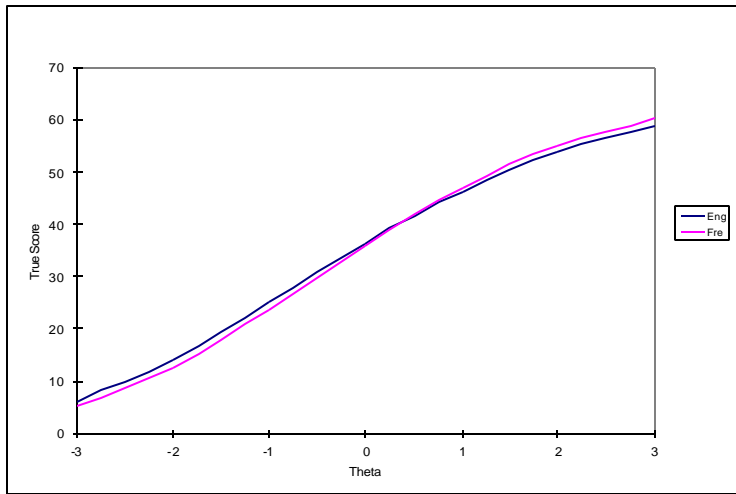
Panel D



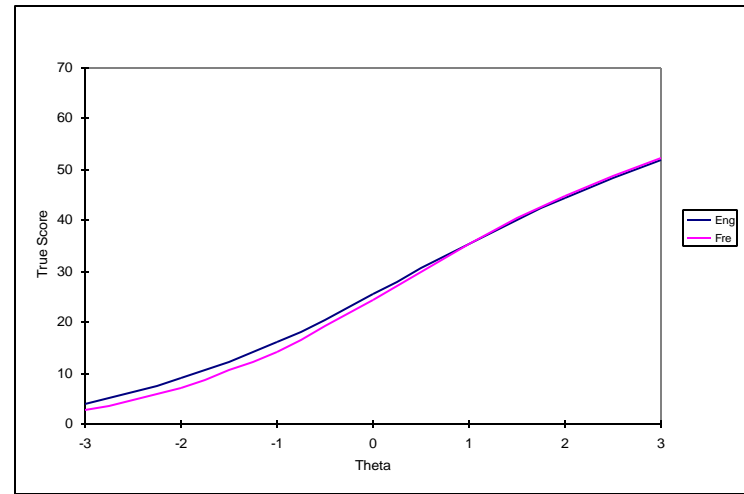
Panel A



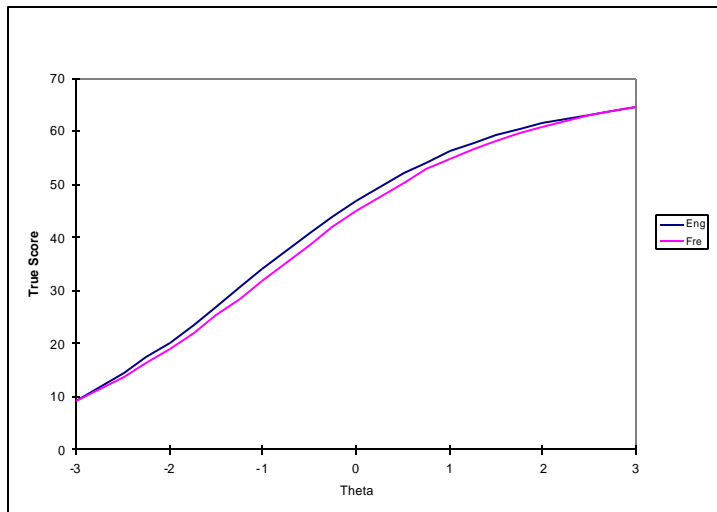
Panel B



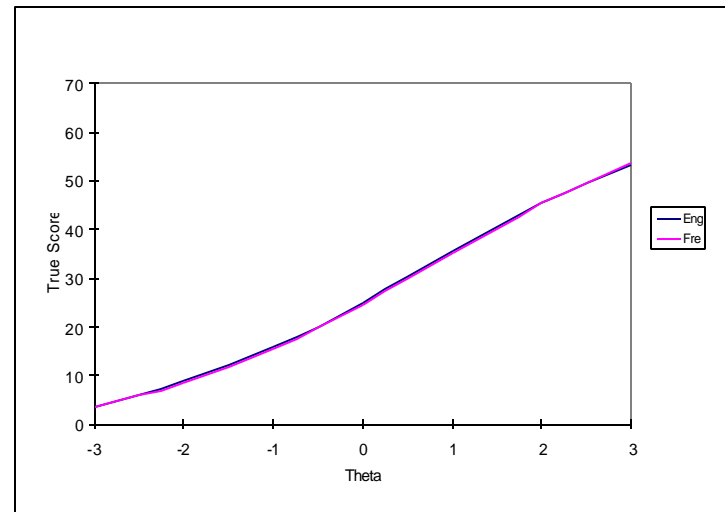
Panel C



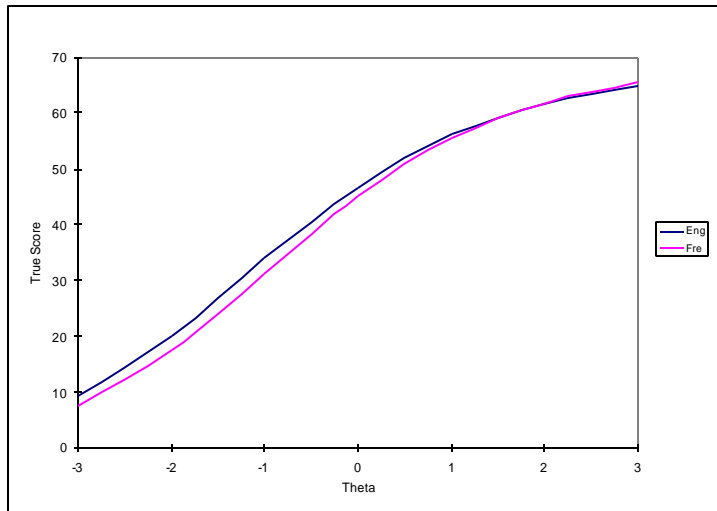
Panel D



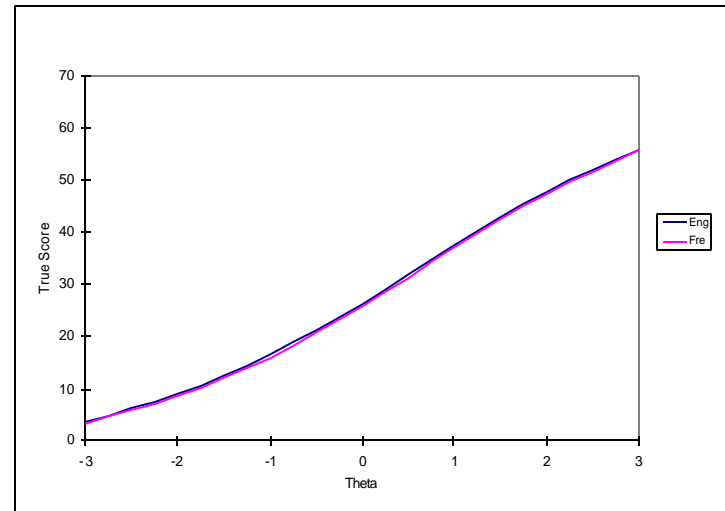
Panel A



Panel B



Panel C



Panel D