Please cite as: Roth, W.-M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating sources of differential item functioning using expert think-aloud protocols. *International Journal of Science Education, 35*, 546–576.

Investigating Linguistic Sources of Differential Item Functioning Using Expert Think-Aloud Protocols in Science Achievement Tests

Wolff-Michael Roth, Maria Elena Oliveri[2], Debra Dallie Sandilands[2], Juliette Lyons-Thomas[2], & Kadriye Ercikan[2]
[1]*Griffith Institute of Educational Research,* [2]*University of British Columbia*

ABSTRACT Even if national and international assessments are designed to be comparable, subsequent psychometric analyses often reveal differential item functioning (DIF). Central to achieving comparability is to examine the presence of DIF, and if DIF is found, to investigate its sources to ensure differentially functioning items do not lead to bias. In this study, sources of DIF were examined using think-aloud protocols. The think-aloud protocols of expert reviewers were conducted for comparing the English and French versions of 40 items previously identified as DIF ($N = 20$) and non-DIF ($N = 20$). Three highly trained and experienced experts in verifying and accepting/rejecting multi-lingual versions of curriculum and testing materials for government purposes participated in this study. Although there is a considerable amount of agreement in the identification of differentially functioning items, experts do not consistently identify and distinguish DIF and non-DIF items. Our analyses of the think-aloud protocols identified particular linguistic, general pedagogical, content-related, and cognitive factors related to sources of DIF. Implications are provided for the process of arriving at the identification of DIF prior to the actual administration of tests at national and international levels.

Key words: think-aloud protocols; experts; translation; item [language] differences; performance differences;

INTRODUCTION

Science educators have shown great interest in international testing programs such as the Programme for International Student Assessment (PISA) and Trends in Mathematics and Science Study (TIMSS), as indicated by related special issues in both the *Journal of Research in Science Teaching* (Bybee, Fensham and Laurie, 2009) and in the *International Journal of Science Education* (Olsen, Prenzel and Martin, 2011). Educators and policymakers use such results to compare their students "to a common standard and to compete among themselves for top scores" (DeBoer, 2011, p. 567). This may, especially when the students of a country score below expectations, lead to shock and soul searching. Thus, in Germany, low results have led to recurrent shocks felt by the entire society, leading the German weekly *Der Spiegel* to publish a "Review of a Decade of Shock." But is it possible to compare the achievements of students from different countries, especially if they take tests in different languages; and are there regional differences in those nations where there are multiple national and test languages, such as Canada (French, English), Belgium (Flemish, French), Switzerland (German, French, Italian, and Romansh), or Finland (Finnish, Swedish)?

Comparisons of results of multi-lingual international and national tests of science achievement are employed to make significant resource allocation, curriculum planning, and strategic decisions. The validity of inferences based on scores of multi-lingual assessment tests, however, critically depends on the comparability of scores. That is, it is vital to determine whether the assessments measure what they are designed to assess rather than construct-irrelevant factors such as the quality of test adaptation including differences in vocabulary, grammar, or semantic nuances (Arffman, 2010; Grisay & Monseur, 2007). Even though test developers spend considerable resources on developing comparable versions of multi-lingual assessments, subsequent psychometric analyses often exhibit *differential item functioning* (DIF) (Artelt & Baumert, 2004). Sources of DIF need to be investigated to minimize bias in multi-lingual large-scale assessments. One important implication from a recent study was that the search for sources of DIF should include information about the translation and adaptation

process, suggesting *think-aloud protocols of those doing the translation work* as a source of direct information about it (Arffman, 2010). The purpose of this study was to investigate potential sources of DIF in a multi-lingual large-scale assessment by conducting think-aloud protocols with experts who are experienced in translating and verifying such translations for a governmental department in one Canadian province. Think-aloud protocols were conducted while the experts compared the English and French versions of 40 previously identified DIF (N = 20) and non-DIF items (N = 20) using data from the School Achievement Indicators Program (SAIP), science domain (Ercikan, Gierl, McCreith, Puhan and Koh, 2004).

BACKGROUND

Large-scale assessments of student learning in mathematics and science – e.g., Pan-Canadian Assessment Program (PCAP), PISA, and TIMSS – play key roles in guiding education policy and practice especially when there are differences between identifiable groups within a population (e.g., gender, class, race, ethnicity, culture). These assessments are conducted at provincial, pan-Canadian, and international levels and the results are used to compare gender and ethnic groups, Anglophone and Francophone students, and countries in the case of international assessments. The validity of comparisons critically depends on whether assessments capture similar knowledge and competencies (construct comparability) and whether the performance results are comparable (score comparability) across comparison groups (Geisinger, 1994; Hambleton, Merenda and Spielberger, 2005). Bias refers to "the extent to which there is evidence of *differential validity* for any relevant subgroup of persons affected" (Bond, Moss and Carr, 1996, p. 17). Differential validity can be due to test content and administration (internal sources of bias), as well as how assessment results are interpreted in relation to some practical or theoretical interest (external sources of bias). Most bias research has focused on internal sources of bias, in particular, evidence regarding construct and score incomparability. Differences in what is measured can make the test more difficult for one of the comparison groups and require competencies/skills not intended to be measured. In both cases, one of the groups may be disadvantaged. In this paper, we focus on linguistic/cultural bias. Any bias is important to

understand, but linguistic/cultural bias has a special pertinence to multicultural societies such as Canada. In particular, any possible bias between English and French versions of tests in Canada could have significant political implications in our nation, which was founded on bilingualism/biculturalism. Unfortunately, there is tremendous evidence of linguistic bias, in particular against Francophone students, on tests that are typically developed in English and adapted to French (Ercikan, 2002).

Bias for Linguistic Groups

Research has shown that in different language versions of tests, differences in signification ("meaning"), key words, and phrases that guide examinee thinking, and varying difficulty levels of vocabulary and sentences can lead to differences in how students read, interpret, and understand test questions. This affects what is assessed and the probability of a correct response to test questions (Hambleton et al., 2005). In French and English versions of Canadian assessments, up to 60% of the items were identified as DIF, and potentially biased (Ercikan, 2002). Possible reasons for measurement incomparability can be derived from language philosophy. Many language philosophers agree on two major aspects of any language: (a) the sense of a word is a function of language as a whole rather than of a single definition; and (b) language and social life are irreducibly interwoven (e.g., Derrida, 1996) so that "the relation of content and language is completely different in original and translation" (Benjamin, 1972, p. 15). An immediate consequence of the interweaving of life and language is that, in some sense, translation from one language to another is impossible (Ricœur, 2004) because each word is part of language as a whole and this language is interwoven with culturally specific life forms more generally. Across two cultures, two languages, two forms of life, the "same word," even as simple as "bee" or "cylinder," can elicit very different semantic relations which affect the trajectory of thought processes (e.g., Ercikan & Roth, 2006).

Methods for Examining Bias

Among the most commonly used methods for examining item comparability are statistical analyses of DIF.[1] A test item is considered to be functioning differentially if examinees of equal ability from different groups have unequal probability of correctly responding to that item (Hambleton, Swaminathan and Rogers, 1991). The presence of DIF indicates differences in the probability of correctly responding to an item and possibly what the test item measures. DIF indicates lack of comparability and possible bias. Whether an item is biased is not solely based on DIF detection. One critical step is the identification of sources of DIF. This step is critical for understanding and deciding upon two key issues: (a) how items affect the validity of interpretations about measurements in different groups and (b) whether to eliminate or revise items to minimize bias. Expert (or judgmental) review – reviews of items by individuals who are knowledgeable about student learning and may have cultural or linguistic expertise – is the most common method for identifying properties of test items (e.g., content, format, context, language) that may cause DIF. However, expert reviews tend to be based on surface characteristics of items (e.g., Ercikan et al., 2010); whether these surface characteristics are the sources of DIF remains an empirical question. Therefore, even though expert reviews can identify whether certain aspects of test items are *associated* with DIF, they may not identify real sources of DIF. Furthermore, expert reviews do not explain *how* specific surface characteristics may lead to differential performance between examinee groups. To answer the questions of *how* and *why*, it is necessary to understand the interaction between the test item *language* and *examinee thought processes*.

Think-aloud Protocols

[1] There are different methods for identifying DIF. The method that had been used in the original study for identifying DIF (Ercikan et al., 2004) is based on item response theory. For each group in a comparison, the difference between predicted and observed mean score is computed for each decile for each item. The observed probability for responding correctly is obtained from the actual examinee responses. The predicted probability of responding correctly is computed based on parameter estimates of the entire sample and the estimates for the members of the focal group. A z statistic is then calculated for each decile and an average Z statistic for the item as a whole. An item is flagged as biased for or against a group when $|Z| > 2.58$.

Think-aloud protocols have been an important tool in understanding and comparison of expert and novice problem solving cognitive processes (Ericsson & Simon, 1993). Think-aloud protocols in domains such as science and mathematics have proven to be ideal tools in investigating how students as well as scientists go about solving problems (e.g., Roth, 2009; Schoenfeld, 1985). Think-aloud protocols do not necessarily ask research participants to talk *about* what they are doing but simply to verbalize what is currently present in their consciousness (Levels 1 and 2 in Ericsson and Simon's, 1993, scheme). That is, participants are encouraged to "think out loud" when completing a task. For this reason, there are no or few mental resources required for producing the think-aloud protocol, which would otherwise take away from the cognitive resources normally spent on the task. Thus, for Level 1 and 2 items, experimental "studies gave no evidence that verbalization changes the course or structure of the thought processes" (Ericsson & Simon, 1993, p. 106).

In the context of identifying sources of DIF in English and French versions of a Canadian assessment, think-aloud protocols have been proposed recently as a new approach (Ercikan et al., 2010). These researchers examined whether examinees' think-aloud protocols confirmed the linguistic differences identified by expert reviewers as sources of DIF. The think-aloud protocols confirmed differences identified by expert reviews for only 10 of 20 DIF items. The low agreement between examinees' think-aloud protocols and expert reviews has two key implications for methods used in investigating bias: (a) item features identified in expert reviews may not be the real sources of DIF; and (b) decisions about item bias cannot be based solely on expert reviews. The Ercikan et al. study only focused on DIF items that were identified to have linguistic differences but did not include non-DIF items or DIF items that did not have linguistic differences.

METHODS

The purpose of this exploratory study was to investigate the identification of sources of DIF by experts involved in translation work as part of their employment. Think-aloud protocols were conducted using a total of 40 science items from the English and French versions of SAIP

originally investigated by another study (Ercikan et al., 2004). The SAIP assessment had 144 science items in total. Of these 49 were identified as DIF in previous research. Expert reviewers identified a total of 17 mathematics and 28 of the science DIF items as having linguistic differences. In the current study, 20 items were randomly selected from the pool of DIF items that were identified as exhibiting linguistic differences by an independent set of bilingual expert reviewers in the Ercikan et al. study, and 20 were randomly selected from the pool with identified non-DIF items.

Translation Experts

Think-aloud protocols were conducted using three participants with expertise in three different areas: (1) pure translation, (2) translation + general pedagogical knowledge, and (3) translation + general pedagogical knowledge + pedagogical content knowledge. These experts had been involved in translation work and verification of English and French versions of provincial curricula and test items for approximately a decade each. (Pseudonyms are used throughout the study.) All three individuals were responsible for verifying the level and appropriateness of the translations of the provincial curriculum and K–12 science assessments prepared by certified translators.

The expert in translation, science pedagogy, and general pedagogy, Paul, speaks French as his mother tongue. He started his career in education more than 40 years ago as a mathematics teacher in Sierra Leone (West Africa). He then worked both in Francophone and Anglophone environments as a teacher, school principal, pedagogical advisor, and translation consultant for K–12 education materials. He obtained his graduate degree in an Anglophone setting. At work, Paul speaks English (predominantly) and French (less dominantly); he speaks French at home.

The translation only expert, Sabina, has considerable experience in English-French translations in educational settings without, however, formal training or experience in science teaching or general pedagogy. Her mother tongue and main language is French. She grew up and lived in France for 29 years. For 26 years she has worked in educational settings, including 7 years as editorial assistant of a journal for French immersion faculty and teachers housed in an

Anglophone university, 8 years as research assistant in French immersion contexts and as project manager for Francophone associations and school districts in minority settings, and 11 years in her current workplace. At the time of study, she was working on verifying the English-to-French translations of Grade 10 science and mathematics provincial examinations. As Paul, Sabina speaks English (predominantly) and French (less dominantly) at work; she speaks French at home.

Walter had immigrated to Canada from the Middle East, with French and English being his second and third languages. He had taught science in French immersion settings and in English to students in grades 8–10, had been a learning assistance teacher, and, most recently, had taught for a year in grade 5. Although he initially felt more comfortable in French, English has become more important over time because of "daily interactions and use, and graduate studies in English," which led to the fact that "English gradually replaced French." He suggested having different areas of linguistic expertise developed through his daily interactions at work in everyday life, such that his food-related language may be most developed in Arabic, and language related to the science curriculum is most developed in English. One might characterize Walter as a science expert whose native language is neither English nor French, though English is his dominant working language for educational contexts. Walter speaks English (predominantly) and French (less dominantly) at work; he speaks Arabic at home.

Procedure – Think-Aloud Protocols

Think-aloud protocols were conducted in three independent sessions, one session with each expert. The three experts were asked to think aloud while evaluating the equivalence of the English and French versions of the 40 SAIP items; 50% were differentially functioning items; eleven had been flagged as favoring English speaking-students, and 9 favored French-speaking students. The experts were not told which items had previously been identified as DIF. Rather, they were told that tests in different language versions sometimes contain items that favor one group over another and the items they were going to review may or may not have been of that type. Prior to reviewing the 40 items, each expert received 3 items to practice the think-aloud

protocol procedures. We printed items on 11" x 17" sheets of white paper containing about 3 items per page. The English version was consistently printed on the left half. The experts were advised that they could mark up these sheets as they wished. Experts reviewed the items in the same order of presentation.

The experts were presented with the same set of written instructions that had been adopted and modified from another study (Ercikan et al., 2004); the instructions were printed on a 8" x 11" sheet of white paper in landscape format and 14 point Times Roman font. The experts were asked to read aloud both English and French versions and then rate each item according to a four-point scale: 0 = No difference between the two versions, 1 = Minimal differences between the two versions, 2 = Clear differences between the two versions, which are NOT expected to cause different performances of two groups of examinees, and 3 = Clear differences between the two versions, which ARE expected to cause different performances of two groups of examinees. The experts were also asked to indicate their levels of confidence on a four-point scale: 0 = not confident, 1 = somewhat confident, 2 = confident, and 3 = very confident. They were also asked to explain any differences they noted and to state the group of examinees (French or English) that they thought the item would favor and why it would do so. The experts were requested to indicate the translation problem and finally they were invited to indicate any additional comments about the item.

Whenever an expert stopped thinking aloud, he or she was immediately asked to continue speaking aloud using prompts such as: "Say what you are thinking," "Read aloud," "What are you thinking right now?," "Would you have translated this differently?," "What do you notice?," or "What are you looking at?" Prompts tend to be required not because the task burdens the experts' mental resources but because the task requirement to think-aloud is unfamiliar (Ericsson & Simon, 1993). Because the experts tended to provide the information requested by the instructions not in the original order, they were requested, when needed, to address an issue not yet addressed. In some instances, experts were asked for clarification right before moving to the subsequent item. For example, Sabina noted during her protocol of Item 3 "I would verify if one

could say ‘surface of the windows’?” Before moving to Item 4, the researcher asked “You say ‘I would verify’?,” upon which she explained that at work she would have verified, using a dictionary, whether one would say in this context “*la taille ou la superficie* [size or surface].”

All think-aloud sessions were recorded with a HD-quality video camera and subsequently transcribed. Paul, Sabina, and Walter spent 144, 98, and 85 minutes, respectively, talking about possible differences between the French and English versions of the 40 items. The lead author, who also collected the data, is fluent in both English and French. The experts therefore were offered the option to conduct the think-aloud protocols in the language of their choice. This alleviates a possible problem: talking in a mother tongue minimizes within-subject translation and “avoid[s] the problem of limited L2 production abilities” (Lee, 1986, p. 204). Paul and Walter conducted the sessions predominantly in English, whereas Sabina predominantly thought aloud in French.

Protocol Analyses

Expert ratings, ranging from 0 (no difference between the two versions) to 3 (clear differences between the two versions that were expected to lead to performance differences) were used to calculate summary statistics. We used standard content-oriented text analyses (Roth, 2005a; Roth & Hsu, 2010; Krippendorff, 2012) to identify categories in the reasons for identified differences between two language versions and in the language problems identified. Content analysis focuses on the content present, involving frequencies of words and concepts and more qualitative accounts of the words and ideas used. In contrast to hermeneutic interpretative methods, content analysis does not attempt to establish “meaning” or the “(essential) truth” of a textual item.

Analysts of think-aloud protocols have been encouraged to take the contents of the protocols at face value and to read them in the context of the situation (Ranking, 1988). Simple summary statistics across experts (which here differed according to their academic background and experience) and items, such as the number of times grammatical structure is mentioned as a translation problem, would have provided an inappropriate and incomplete picture. Therefore we

read the transcriptions purposively to identify themes and issues that repeatedly arose for one expert or between experts. For example, all three experts commented on the fact that the French version tended to be longer (more words) than the English version:

> *C'est beaucoup plus long. Mais ça c'est normal. . . . C'est négligeable. . . . Mais souvent le français est plus long que l'anglais. D'une part*. [It's always longer. But this is normal. . . . It's negligible. . . . But often French is longer than English.] (Sabina)
>
> The flow in English and the time it's taking is, is longer in in French. (Walter)
>
> Here where there's less reading to do. . . . When you translate from English into French, there's always twenty percent more longer, it's always twenty percent longer. (Paul)

We then created the category "length of English/French texts." We read all three protocols to identify all the other instances where an expert might have identified the length of the French versions as a particular issue. Three authors used the final category set to code all three transcripts in their entirety. There was a high level of consistency among the raters. For example, in an initial independent round of coding Paul's transcript, 10 of the 40 items included a category where one rater scored a category alone. In other words, out of a total of 85 coded instances in the 40 items, two or three raters identically coded 69 (81%) of these. In cases where a category was not identified by two or three coders, differences are attributable to the fact that multiple categories were applicable but one or the other coder only chose one of the possibilities. For example, one difference occurred on Item 10 where Paul said: "Type of damage, *l'état* [state] here, *de cette falaise* [of this cliff] slight difference again. Here [French version] we're saying the state of the cliff, here [English version] we're already talking about what kind of state we're talking about damage." Two raters had checked the category "Forms of inclusion" and two raters had checked "cognitive conceptual." Clearly, Paul articulated different levels of inclusion, state versus kind of state; but there also was a cognitive conceptual difference between the expressions "type of damage" (*le type de dommage*) and state (*l'état*). A full appreciation of the differences articulated by the experts at times required great familiarity with French, for example, when Paul points out that *à la maison* (at home) is not the same as "in and around the house." The English describes spatial relations with respect to a physical house, whereas the French term is equivalent

to *chez soi* (Rey, 2012) literally "at one's own/self." Like the English "at home," *à la maison* is used in adverbial phrases. The equivalent to the English "in and around the house" would be *dans la maison et autour de la maison* or *à l'intérieur et a l'extérieur de la maison*. We evaluated all differences to come to a consistent assessment of all occurrences of all coded instances.

We also read the transcripts to determine whether there might be variations in the ways the three experts interpreted such differences. Thus, we found that for Paul and Walter, both former science teachers with Masters degrees, for example, the length (or reading level) of the text had direct implications for the amount of time it would take the student to read and process the texts, inherently disadvantaging the Francophone students because they would have to read more text in the same amount of time as the English students. Sabina, who does not have teaching experience, noted the difference in length without drawing implications for the processing of information or the impact it would have on students' performance.

Under normal circumstances, the dispositions and skills pertaining to a particular practice do not show themselves and are largely invisible. On the other hand, in situations of trouble, where the normal ways of operating in some domain do no longer work, when something appears odd, the dispositions and invisible skills tend to become visible. Thus we expected items where the experts agreed that there were no differences – e.g., most of the non-DIF items – to reveal much less of the reasoning underlying the equivalence.

EXPERT THINK-ALOUD PROTOCOLS – SUMMARY STATISTICS

In this study, sources of differential item functioning (DIF) were investigated by using experts with more than a decade of experience in the verification of French translations of English curricula and examinations. The participants differed in background and degree of linguistic and pedagogical expertise. We present our results in two parts. First we provide summary statistics of the experts' ratings of the items according to the 4-point scale described above along three dimensions: (a) is language expertise sufficient to identify DIF items?, (b) are language experts consistent in their ratings of DIF and non-DIF items?, and (c) is there a bias in

identifying the direction of possible DIF? In the second findings section we describe and explain results of the think-aloud protocols along five major possible sources of DIF that emerged from our protocol analysis.

Is Language Expertise Sufficient for Identifying Items?

The results show that even experts with a lot of experience in the verification of English-French translation of official curriculum guidelines and provincial examination do not a priori identify those items that have shown to lead to DIF (Table 1). Even among those items that did not lead to DIF, the experts sometimes identified differences, some of which might lead to lower performances of Francophone students, especially in minority settings – where students often fall back on English as soon as the teachers are out of range (Roth, 2005b). In all cases where an expert identified possible performance differences (n = 13), these were to be in favor of those students taking the test in English – an assessment that might have a possible reason in the fact that the experts are responsible for French language programming in a non-French speaking province. However, the different evaluations of language problems provided might well be associated with the background and experiences of the experts.

«««««« Insert Table 1 about here »»»»»»

Paul, native French speaker and expert in general and specific pedagogy, identified 5 DIF items (25%) as being linguistically different, all of which he suggested to lead to performance differences though in 4 cases (20%), the direction of the advantage was in the reverse direction (favoring English instead of French test takers). Of those 15 DIF items (70%] that he identified as exhibiting no or minimal differences, the majority (n = 14 [70%]) should not lead to performance differences; one (n = 1 [5%]) item might lead to performance differences for Francophone students in minority settings (i.e., not in Quebec or New Brunswick). Of the 20 non-DIF items, a large number (n = 14 [70%]) exhibited no relevant differences, though n = 3 items (15%) might lead to performance differences among Francophone students in minority settings. In two instances, he identified non-DIF items as having the potential for leading to

performance differences, whereas 1 item (5%) was recognized as being different without leading to performance differences.

Sabina, with language but no pedagogical expertise, identified most DIF items (n = 19 [95%]) as exhibiting no or minimal differences, one (n = 1 [5%]) DIF item as different without leading to performance differences. She classified n = 19 non-DIF items (95%) as exhibiting no or minor differences and 1 item (5%) as being different without leading to performance differences.

Walter, with expertise in science pedagogy but French/English as second/third languages, identified 15 (75%) of the 20 DIF items as exhibiting no or minimal differences. Of the five (n = 5 [25%]) items that were identified as exhibiting differences, n = 3 (15%) were to lead to performance differences whereas two (n = 2 [10%]) were not expected to cause differences in performance. A majority of the non-DIF items (n = 14 [70%]) was potentially leading to differences; and Walter suggested that n = 3 non-DIF items (15%) might lead to performance differences, whereas another n = 3 non-DIF items (15%) were said to exhibit clear differences between the two versions without causing performance differences.

Are Language Experts Consistent Among Each Other?

The answer to this question is summarized in Table 2. The three experts agreed on 10 (50%) of the DIF items as exhibiting minimal or no difference; on four (n = 4 [20%]) of the DIF items, two experts agreed that these are different (with or without performances difference) with the third suggesting no/minimal difference; on 6 DIF items (40%) one expert suggested there is a difference and two that there were no/minor differences. The three experts did not reach agreement on any one item (n = 0 [0%]) as being clearly different. That is, the three experts were unanimous in their assessment of 10 items (45%), with the two experts with science teaching experience agreeing on 12 (60%) of the items, whereas one or the other recognized the remaining item as different (with or without causing performance differences) whereas the respective other identified no or minimal differences.

«««««« Insert Table 2 about here »»»»»»

Of the 20 non-DIF items, the three experts reached complete agreement about the equivalence of 11 (n = 55%) items as being not different. On 8 non-DIF items, at least two experts agreed that there are no or only minimal difference between the English and French versions. One item (n = 1 [5%]) was identified by all three experts as being different with two suggesting performance differences whereas one suggesting no performance differences. That is, all three experts were in agreement on 12 (60%) of the items about there being no/minimal or clear differences between the two language versions (first row). We did not find it useful to compare these outcomes to random performance – such as what one observes when students pick an answer on a multiple choice item when they do not know how to respond – because the experts based all of their choices on the specific aspects that they noted while comparing the two versions.

Is There a Language-Related Bias in Evaluating Linguistic Difference?

Because all three experts either spoke French as their native language (Paul, Sabina) or learned French before English (Walter), the question arose whether there might have been a bias in the assessments of DIF items favoring English versus those that favor French. (All three had been reminded that the bias could be either way.) On the 11 DIF items favoring English, there were 3 items where two experts agreed that there were *clear* differences between the versions, and on two of which performance differences were anticipated (Table 3). One expert, who anticipated performance differences, identified another item as favoring the English version. On the 9 items that had favored French test takers, only one expert identified it as clearly different, leading in 2 cases to performance differences. In sum, therefore, there was a higher proportion of DIF items identified by two experts when they disadvantaged French speakers (DIF_F) than when they advantaged English speakers (DIF_E).

«««««« Insert Table 3 about here »»»»»»

POTENTIAL SOURCES OF DIFFERENTIAL ITEM FUNCTIONING REVEALED THROUGH EXPERT THINK-ALOUD PROTOCOLS

As noted above, our analyses of the think-aloud protocol transcriptions using standard methods of content analysis identified five major categories as potential sources of DIF between the two language versions. These are differences in (1) relative length of the two language versions, (2) linguistic issues, (3) logical structure in content or form of the item, (4) cognitive-conceptual content, and (5) diversity issues. An overview of the categories, descriptions, examples from the transcript, and our code are provided in Table 4. In the protocol analyses, we provide word-for-word translation of the French items and quotations of experts. We denote whether the featured and discussed items had been identified as DIF items (DIF_E, favoring English, DIF_F, favoring French, and nDIF, non-DIF item). Each category is described and exemplified in the following sections.

«««««« Insert Table 4 about here »»»»»»

Length of the Different Language Versions

All three experts repeatedly noted that the French version of many items was longer than the English version. Paul tended to quantify length differences, whereas Sabina and Walter did not. All experts stated that the additional length of the items would demand more time from French-speaking examinees. Although the three experts suggested various ways to shorten the text, Sabina noted that in many cases French is longer by nature: "Often the French is longer than English, for one" (Sabina, 16, DIF_F). Paul but not Walter or Sabina suggested that this could lead to differences in: the amount of (a) reading, (b) focus and mental energy required from French test-takers, and, therefore, that this would ultimately affect their performance.

To illustrate, Paul pointed out that the distractors in the English items were written using only one-word (e.g., herbivore, carnivore, scavenger, decomposer) whereas the French version used phrases (e.g., *Ce sont des herbivores* or *Ce sont des carnivores*). Paul asked, gesturing toward the first part of the phrases, "*Ce sont des*, Why add this?" He suggested that the English version of the assessment used one-word options and queried why the extra words were added to the French version, noting that the English students could quickly browse over the answers whereas the French students "have to do all the reading."

The experts also noted in various places that the French contained additional text that provided more specific information. For example, Sabina noted in Item 31 (DIF_E) that the English version said "this investigation," and the French version further specified: "*son expérience sur les lotions solaires* [her experiment of the sun lotions]."

Paul consistently associated the length of the text with the cognitive demands on the students taking the test. Thus, he suggested,

> the major difference if you exclude the context you know with the kids (inaudible) schools, family, all that . . . so ah . . . it is twenty percent longer right, so it's twenty percent more reading time, more comprehension right, so that's . . . and so I think it's a factor that could impact on the kids. (Paul, 40, DIF_F)

He also made reference to the fact that this makes demands on the brain, "You see always twenty percent longer, more brain power" (Paul, 23, nDIF). For equivalent grade 10 students taking the same exam, "the pressure on the French kid would be much more, much higher. I mean the length factor, the one hour and a half time factor, comprehension" (Paul, 39, DIF_F).

As Paul, Walter emphasized the cumulative effect on performance that would favor the English over the French test taker: "longer time the only items I said longer time will give you a sense at the end why an item like this might be more favorable for the English-speaking students" (Walter, 10, DIF_F). Paul elaborated on the cumulative effect by describing his own experience of being more tired at the end of the day even though he has lived 12 or 15 years "in an English set up."

In summary, all three experts emphasized the considerable length differences between the English and French versions. The two experts with pedagogical experiences (having worked as teachers) tended to highlight the amount of processing time certain items would take – because the French text is longer or because the French is more difficult to understand, of a higher complexity, than the English version. Although the language-only expert provided examples for making a text shorter, she also pointed out the general tendency of French to be longer.

Linguistic Differences

Among the linguistic differences, experts noted (a) higher general reading level of French items, (b) the use of unfamiliar, often more literary words and expressions, (c) grammatical differences (grammar, tense), and (d) semantic differences deriving from differences in figure–ground relations.

General Reading Level The experts repeatedly noted that the French version had more complicated sentences that increase the reading level. They suggested that the text could have been expressed more simply to make it closer to the English version (e.g., "So there could have been a much simpler sentence structure," Paul, 16, DIF_F). The experts also pointed out that they sometimes had to read the French version twice before understanding it, whereas it took them a single reading to understand the English version. For example, for Item 2 (DIF_F), Walter points out that the reading level is not associated with a specific word, but an overall effect resulting in an increased demand on time. Walter also pointed out specific elements in the French version that made it more difficult. For example, he stated: "In English, 'a group of students are participating,' very straightforward. I read it in French, *Un groupe d'élèves participe,* and I'm sure, it's not just my brain, some students might take a few seconds to think, 'Why is it *élèves aux plurielle* [students in plural] and *participe* [participate] is not *ent*?" (Walter, 5, DIF_E).

He also related the reading level to the length of processing needed by the French-speakers: "The level this is stated would take slightly longer for the French side to know *debordement* is spilling *nuirait au développement* [would be harmful to development]" (Walter, 39, DIF_F). He concluded, "takes a little bit longer in terms of assessing the information reading the item more than once, which should take longer" (Walter, 39, DIF_F). Similarly, Paul, exemplified the higher reading levels with specific examples:

> And here is quite a long *Parmi les caractéristiques suivantes, laquelle ne se retrouve pas chez la plupart des êtres vivants* [Among the following characteristics, which is not found among most living beings]. I find the English here– it's easier to understand quickly, but the question is about which is **not** characteristic of the thing right. Here, even myself, I had to read twice the sentence to make sure what they wanted here. (Paul, 2, DIF_F)

Sabina makes the same point in the context of Item 6 (DIF_E), "Which of these are the result of a process that can be reversed?" (*Lesquels sont le résultat de changements réversibles, qui peuvent être inversés*?). She suggested that the question is not very clear in French, then articulated its parts comparing these to English and concluded these to be the same. She commented, "It's because it's always simpler, the English version often is simpler than the French version . . . It's clearer." She added, "It would take me some time to try and change this, to see whether there is a better way of expressing this turn of phrase."

There is only one instance where one of the experts noted the French to be less complex. In Item 7 (nDIF), the English version "Galileo stated that Earth orbits the Sun, rather than that the Sun orbits Earth" repeats part of the phrase, whereas the French version *Galilée affirmait que la Terre décrit une orbite autour du Soleil et non l'inverse* [Galileo affirmed that Earth describes an orbit around the sun and not the inverse] uses the shorter expression "and not the inverse." Sabina stated, "For once, the reverse has happened . . . here they reduced the French version with respect to the English version."

Unfamiliar Words and Expressions, and Semantic Differences The experts pointed out particular words and expressions in the French version that are more infrequently used in French as compared to the equivalent word used in the English version. Such words tend to be "not a preference [in French]" (Paul), even among French speaking people from Quebec. In Item 26 (DIF_F), for example, the English version uses "artificial hand," which in French would be *main artificielle*. However, the term used in the French version is *prothèse*. Sabina reasoned that the English version constitutes a precision, whereas *prothèse* is a more general (science) concept. She would have verified the translation, as she noted the existence of the English term prosthesis, which might have been an alternative. Walter, too, noted a clear difference between the two expressions but marked it as moderate (without causing performance differences). He suggested that at the examinees' age level, more Anglophone students would be familiar with "artificial hand" than Francophone students would be with *prothèse.*

Sabina suggested that there was a (semantic) difference between the English "possible food source for consumers" and *source possible d'alimentation des êtres vivants* [possible source of food of living beings]. She suggested that *êtres vivants* implied humans or animals, and she did not know whether the same was the case for "consumer." She then inferred that *êtres vivants* consume [*consommer*], so that this would constitute a negligible difference between the two versions. Walter highlighted differences in particular words:

> It is again the the level of of language use – lightning has struck very clear for an English speaking students, *la falaise a été foudroyée par un éclair* [the cliff has been struck] how many student would be familiar with *foudroyée* [struck] I am, I am not sure many of them will guess it. (Walter, 10, nDIF)

The French version, he noted, indicates a state, whereas the English version identifies damages specifically. Paul made exactly the same point and highlighted that in one case you know that there is damage whereas the French version only talks about its state. The two also agreed on another term to be unusual. Both pointed out that the French expression *a été foudroyée* [has been struck] tends to be used with persons rather than with inanimate objects and Paul proposed that the French ought to be *la foudre est tombée sur la falaise* [The lightning fell on the cliff]. Similarly, Walter expected English students to be familiar with the term "cattail," whereas he anticipated less familiarity with the French equivalent *quenouille*.

There were many other examples of translations where the French version makes use of less common words than the English version. The experts included the instance where the English term "skin color" was translated by the term *complexion* [complexion], where the French offers a direct equivalent, *couleur de la peau*, a choice actually made in another item of the same test.

> So here the use for skin color they use the word complexion, but complexion is English right, English word, skin complexion. I would have used skin, *la couleur de la peau* [the color of the skin]. *La même complexion* [the same complexion]*, uh* . . . is not a very common word. It certainly is a word in French of course. (Paul, 30, nDIF)

Sabina, too, suggested an alternative to the translation of skin type (*le teint*) and then produces some context for evaluating the alternatives ("it's to situate whether you have light skin or dark skin [*la peau claire ou la peau foncée*]"). In another instance, for example, Paul suggested for

Item 5 (DIF_E) replacing the verb *recueillir* [collect] by *ramasser* [gather] because even among Francophone speakers, the former is much more uncommon than the latter.

Paul, as Sabina, pointed out that the French word choice is frequently more literary. Both talked about the magazine published by the national airline *Air Canada*, which has "the best translations . . . they translate using English or in French using French expressions that are equivalent but not the translation of the English expression." Both emphasized that the high quality of the translations present in the airline magazine derives from the fact that English expressions are translated by French expressions that have no word-for-word equivalence. But Paul suggested that the point of a good translation in education is not to produce the best literary text: "You're not in a magazine." He said this should be the case especially for non-Francophone individuals, "This is not something you do in Francophone education outside of Quebec you know." Sabina noted the different requirements to express some state of affairs in French, which leads to a translation that is not word-for-word. For example, Item 6 (DIF_E) states that "The students travel by boat along the coast of New Brunswick," whereas the French version reads *Pendant une excursion en mer, les élèves longent la côte du Nouveau-Brunswick* [During an excursion on the ocean, the students travel along the coast of New Brunswick]. She noted that in French one does not say *qu'ils sont en bateau* [they are in the boat], because there is no other way to go to sea than in a boat. In Item 11 (DIF_E), she noted that the English "hangs them on a long piece of wood to dry" was rendered as *les laisse sécher sur un long morceau de bois* [them lets dry over a long piece of wood]. She suggested that the fact that the "hanging" of the fish was not specified would not penalize French test takers.

Grammatical Differences Experts noted many instances where the French and English versions are not quite the same, including syntactic differences that lead to semantic differences, even though they ultimately judged the differences to be minimal. That is, differences are based on the fact that the grammatical structure of French items frequently was more complex. Relative pronouns, clauses, adjectival contractions, and direct quotation of preceding text tend (have) to

be used differently in the two languages. As a result, grammatical differences lead to longer or more complex French texts making it more difficult to grasp its science content. For example, Paul identified a particular construction using the relative pronouns, which he identifies as having a more complex structure

> "She needs" . . . *voici ce dont elle a besoin*. Why not say *elle a besoin de* [she has a need of]? Why say *voici ce dont elle a besoin* [here is that which she is in need of], right, so I would have scratched that. And *voici ce dont elle a besoin* is a complex structure in French. The use of *don't* is always its who you know a relative, a *pronom relatif* [relative pronoun]. It's a very complex use, it's a more complex, the relative pronoun to use. "Of which," you know, "whose," "of which," you know, "Mary whose dog is sick," so that's the *et voici ce dont* [and here is that which]. (Paul, 30, nDIF)

He concluded, "not every kid would necessarily understand that." Similarly, Sabina identifies grammatical structures that are not only more complex but also render the text longer. For example, whereas the English question "With respect to the volunteers, what is *not* necessarily Helen's responsibility?" contains 11 words and takes up 1.25 lines of text, the French translation *Laquelle parmi les activités suivantes n'est pas nécessairement la responsabilité d'Hélène envers les volontaires?* [Which among the activities not is necessarily the responsibility of Helen towards the volunteers?] contains 15 words and takes up 1.75 lines. In this situation, the test developers shortened the English version with "which among the following activities" to "what" and dropped the indirect object. Sabina suggested that in French "One is obliged to make a longer sentences to specify why "leading to longer text, there is no difference . . . it's because it is longer" (Sabina, 32, nDIF). The same form of construction was noted between the English "Which of these materials can be recycled?" and *Parmi ces matières, lesquelles sont recyclables?* [Among these materials/substances, which are recyclable?]. Again, the French version uses a clause whereas the English version uses the interrogative "which" in pronominal form, thereby introducing with one word what the French introduces by means of a clause.

The English language affords contractions that tend to be impossible in French. Instead, French requires the use of different grammatical constructions, such as a clause. When a

contraction is actually made, then the French often does not sound right, and may become confusing. Paul provided an example.

> So *le mur du sous-sol est construit de quinze centimètres d'épaisseur de briques . . . quinze centimètres d'épaisseur de briques* [the wall of the basement is constructed of fifteen centimeter thickness of bricks] . . . that's a very complex way to say fifteen-centimeter brick wall. *Le mur du sous-sol est construit de quinze centimètres d'épaisseur de briques* . . . that's a very bad translation. The basement wall is made of fifteen-centimeter brick . . . *est construit* . . . Very bad translation, very confusing. (Paul, 20, DIF_E)

Here, the French employs a construction similar to English, where the predicate "is 15 cm thick" that defines the wall is changed and employed in adjectival form "15 cm wall." In French, such a construction generally is not possible, and, in situations such as this one where it is used, it is thought of as constituting a "very bad translation." Paul offered a different construction instead: "So there should be here *est construit de briques de* fifteen uh *de quinze centimètres d'épaisseur* [is constructed of bricks . . . of fifteen centimeters of thickness]." Walter also provided an example where using a contraction leads to a more complex French structure:

> The English "greatest impact" it's "the greatest impact on lowering energy use," it's clear for that age group. Then *la plus forte réduction d'énergie utilisée* [the highest reduction of energy used] uh *plus forte réduction* and the sort of the contradiction between *forte* [great/strong] and *réduction* might be confusing for some students. So again I think the English students are favored by clearly isolating "greatest impact" from . . . sort of "lowering energy" use whereas the French translation jump directly . . . "the greatest reduction in energy use" might lead to some confusion for some students. (Walter, 4, DIF_E)

He pointed out that the contraction in French created the possibility for confusion, whereas the English separated the two tendencies, attributing the first to the impact and the second one to energy use. In this instance, not noted by the experts, the English question actually repeats (direct quotation) the stem ("ways to *lower energy* use," "on *lowering energy* use") whereas the French version changes expressions, thereby using indirect speech (*diminuer la quantité d'énergie* [decrease the quantity of energy], *la plus forte réduction d'énergie utilisée* [the greatest reduction in energy used]).

Experts also referred to differences in tense. For example, Item 34 (nDIF) reads "When Helen has collected her data, what should she do," which is translated as *Lorsque Hélène aura recueilli ses données, que doit-elle faire* [After Helen will have collected her data, what does she have to do]. Sabina and Paul suggested the future and present tenses were used in French whereas in English the *passé [composé]* (present perfect simple) tense was used. Paul clearly marked the grammatical error: "The use here *aura recueilli* [will have collected], right, was a future anterior, *future antérieur,* and then they should use this future anterior then that should use future here, *que devra-t-elle faire* [what will she have to do]" (Paul, 34, nDIF). Sabina noted that "has collected, it's not 'will collect'" and that "what should, *que devrait-elle*" "is not the present tense." In another instance, Item 40 (DIF_F), Sabina noted a clear (grammatical) difference between the English "A blood sample is taken from Jill's finger to test for sugar content" and the French *Avec une goutte de son sang, il serait possible de mesurer son taux de sucre* [With a drop of her blood, it would be possible to measure her sugar level]. In the English, she described, it says that one takes and one tests [*vérifie*], whereas the French uses *il serait possible* [it would be possible], thereby using the conditional. Because the question is about the magnifying glass, however, she did not think the clear difference to lead to performance differences. Paul also noted a different in tense in Item 13 (nDIF), where the English version states "they may not be able to tell if their dog were sick," which he suggested "means be aware that their dog is sick, whereas the French expression *que leur chien est malade* [that their dog is sick], he repeatedly emphasized that "the dog is sick."

Different Figure–Ground Relations Experts also noted the different figure–ground relations that different language versions require – indicating the presence of conceptual differences. For example, in the sub-section on unfamiliar words and expressions, and semantic differences the examples show that the two languages appear to require different ways of articulating the relation between an action and the context. Thus, Sabina suggested that the French Item 6 (nDIF) does not need to state that students travel by boat, because *excursion en mer* [sea excursion]

implies the boat; the French implies that the fish in Item 11 (DIF_E) are hung when the drying process occurs *sur un long morceau de bois* [over a long piece of wood]. Of course, without actual think-aloud protocols of students taking such items, we cannot know whether differences in figure–ground relations lead to differences in reasoning and, thus, performance.

Sabina tended to notice such figure–ground differences more frequently than Paul or Walter – often arising from grammatical structure – that lead to conceptual differences. In fact, the different grammatical structure was required because of the different ways in which English and French relate the focal phenomenon to its context.

> Here there is a little difference, because they say *dans les travaux à la maison* [in the works at home], but here it is "in and around." This, this means, for me this means on the inside of the house, but they don't really speak about the outside of the house. So there is a little difference "in and around the house," but its minimal . . . to be correct, I would add the translation here, *à l'intérieur* [on the inside], I would be more precise, I would say *à l'intérieur et à l'extérieur de la maison* [at the inside and outside of the house]. (Sabina, 1, DIF_F)

Paul made precisely the same point, stating that *à la maison* is equivalent to "at home" but not to "around the house." Both he and Sabina suggested the exact same alternative translation. In this way, all three experts repeatedly suggested that they would have asked for different translations if they faced the situation at work and, as in this case, sometimes proposed precisely that. The alternatives they offered might have led to a higher level of equivalence between the English and French versions.

Sabina also noted differences in the two versions of Item 10 (nDIF) with respect to the fact that an agent of change is indicated in one distracter: "'The rocks at the top of the cliff were blown down.' Here [French] they say that it is the Wind, but here [English] they don't specify that it is the wind." She elaborated, "'Blown down,' does this mean that it is the wind that makes [the thing] fall?" She continued by stating that the English version suggests the wind to be the agent of the change, whereas the French version does not propose but only implies an agent.

Sabina usually rated such differences as minimal (*C'est minim* [it's minim]). Paul often noted the same problems, but then concluded – based on some pedagogical or cognitive principle such

as length of processing time, fatigue, performance level, "learner styles," "reading styles" – that these might lead to performance differences. Whereas the two experts (Paul and Sabina) noted that differences in French expressions might have led to differences in conceptual understanding, Walter, whose native language is neither French nor English, did not comment on the same differences.

Differences in Logical Structure

Experts noted three differences related to logical structure of the two language versions: use of different forms of (a) negation and double negatives, (b) inclusion (general vs. specific), (c) logical relations, or (d) probability.

Negations Paul in particular repeatedly noted occurrences of negative forms, especially double negatives when a term that implies a tendency in one direction (lower, higher) is negated by another. One of the cases pertained to the reduction of heat flow by means of insulation:

> So the difference I see here, when you use *la plus forte réduction,* reduction is this, *la plus forte* is this right. The strongest reduction right . . . so right there you have to think twice. What do you mean by that? Here they talk about the "greatest impact" right, so why not use *le plus grand effet* instead of the greatest reduction. Even myself, I'm sixty-three, I've master degree and go *la plus forte réduction.* I had to read twice. (Paul, 4, DIF_E)

The French version uses the proposition *sa résistance à la perte de la chaleur* [its resistance to the loss of heat], which makes a negative force (*résistance* [resistance]) operate upon a negative process (*perte de chaleur* [loss of heat]), whereas the English version asks for the "value of resistance to heat flow," suggesting a negative force on a neutral process. Such double negation "takes two seconds more and more wearing of the brain" (Paul). Another instance concerned floating and sinking:

> "What kind of force keeps the canoe afloat," *Comment s'appelle la force qui empêche les canots de couler?* [What is the name of the force that prevents a canoe from sinking?] So what kind of force keeps the canoe afloat. What kind of force prevents the canoe from sinking. Okay. So again its different right. [. . .] It's a different brain exercise because it is by the negative right? What prevents from sinking? Here what keeps them afloat. So anyway. I would say, you know, *qu'est-ce qui permet aux canots de flotter* [what allows the canoe to float], very simple. So why – it's like the other questions being in there, right – you have a double kind double negative. (Paul, 25, DIF_E)

Paul provided a specific example of the double negation that could have been used to make the French version more equivalent to the English version, which does not use the double negation and therefore is "very simple." Moreover, he suggested here that the problem occurs elsewhere in the set of items, "the other items in there." He also noted the double negative in the French construction *une orbite autour du Soleil et non l'inverse* [an orbit around Sun and not the inverse], where the English "Earth orbits the Sun, rather than that the Sun orbits the Earth" avoids this double negation. Paul frequently was perplexed, repeatedly expressed as "Holy smokes!" Thus, after reading Item 13 (non-DIF), Paul noted that he had to read it twice to understand what it was about: "the question here means implies that he is not definitely sick right but the statement here implies that he is sick so I have to read twice what they mean by that." In English, it asks "Which information about their dog would *not* help them to tell if it were sick?" whereas the French version states that the dog is sick and asks, using a negation of a positive *ne* les aiderait *pas* [would not help them].

Paul noted an analogous difference in an item where the English version is less specific with respect to the entity that develops ("Maria and Raphael collect a sample of frog eggs to study frog development"), where the development could refer to a frog in general or to the eggs in particular. The French version *Maria et Raphaël recueillent quelques œufs de grenouille afin d'en étudier le développement* [Maria and Rafael collect some eggs of frogs to study (the) development], however, specifically asks for the development of the eggs using the reflexive construction *d'en* [of these]. Although the 40 items included more such cases, not all of the possible candidates for such effects were pointed out.

Categorical Differences Imply Different Levels of Inclusion The experts repeatedly highlighted differences that can be classified as those between more generic (more inclusive) and more specific (less inclusive) terms. The difference is one between category [concept] and case. In French versions, there was a tendency for more general terms whereas more specific terms appeared in the English translations. Thus, Sabina noted that the French version of Item 26

(DIF_F) uses the term *prothèse* [prosthesis], whereas the English version uses "artificial hand," which is a specific type of prosthesis. Several category and type differences were noted in Item 18 (DIF_E), the item for which the three experts generated the most possibilities for differences between English and French items. The item shows a drawing with a farm, a cloud from which it rains, mountains, a windmill, a tree, and the sun. The question reads: "What is one object in this view of the farm that could produce energy to run machines?" (*Dans cette illustration, quel élément peut produire l'énergie nécessaires au fonctionnement de certains instrument de la ferme?* [In this illustration, which element can produce the energy necessary for the functioning of certain instruments?]).

Sabina noted that the term *ferme* [farm] is more inclusive than the term house, which is included as one of the structures on a farm. She also pointed out that "machines" and "instruments" are not the same – but concluded that the end result would be the same, because something, in this case the windmill, produces energy for something. Walter, who noted a clear difference that should lead to performance differences, pointed to the difference between "object" and *élément* [element]: "Um object in English is very clear it's an object . . . nonliving *élément* in French might lend some students to start thinking outside of objects and start sort of venturing into perhaps . . . what are other sources . . . object is very specific than of object is more specific than *élément* so it is a large difference" (Walter, 18, DIF_F). He articulated the different levels of inclusion that are implied in the English object versus the term *élément* [element] that was used as the French equivalent. That is, English version asks for an "object" "in this view of the farm," whereas the French version asks *quel élément* [which element] *dans cette illustration* [in this illustration]. Thus, whereas the English text orients students to the farm and to an object, which Sabina identified to be the windmill, the French version directs students to the illustration and to one of its elements. As the type of windmill seen historically has been used to draw ground water, the French version more easily allows *the sun* as the answer, whereas the English version distracts students away from the sun, which is not generally treated as an "object." However, in the "illustration," the depiction of the sun is indeed one of the elements.

The term *élément* [element] includes the sun (logically inclusion), whereas "object" does not. The effect that such differences bring about is similar to the one noted in the case of figure–ground differences in that the concept used includes or excludes context and orients students differently to the situation as a whole.

Logical Relations between Parts of an Object Experts noted that the two language versions differed in relation to the degree to which different aspects of an object were explicitly related. Thus, for example, *sa résistance à la perte de la chaleur* [its resistance to the loss of heat] (20, DIF_E), uses a personal pronoun *sa* [its] to refer back to the wall of the basement (*mur du sous-sol*) in the statement that directly precedes the phrase, whereas the English version "What is the value of resistance," does not tie resistance to the basement wall. Rather, it refers to value in general. Paul noted that he would expect grade 11 students to know that resistance does indeed refer back to the basement wall.

Sabina (40, DIF_F) noted that the English version stated that "A blood sample is taken from Jill's finger to test for sugar content," that is, that it makes a statement as to what has happened (taking the blood) and what to do with it. In French, she noted, the presence of a conditional, *il serait possible* [it would be possible]. Although she characterized the two versions as "clearly different," she did not anticipate performance differences because the question was about the magnifying glass rather than about the taking of the blood.

The two language versions often used different expressions when in fact the same phenomenon is addressed. Thus, Sabina noted that the English version of Item 20 (DIF_E) talks about "their basement" in the context of the previously stated "farm" (Item 19, DIF_E) whereas the French version uses *sous-sol de la maison* [basement of the house], a term denoting part of a farm and, thereby, introduces a different term than that stated in the context. That is, whereas the second of these items relates *directly* to the preceding item in the English version, the French version uses an indirect relation that requires additional cognitive processing (house [part] → farm [whole]). Sabina also noted the different way in which an item orients students when the

English version suggested "connect each of the three birds shown below" (Item 28, nDIF), thereby making explicit the relation between the text where the threeness of the birds, and the relation to the drawings ("shown below") are explicitly articulated, whereas the French version simply states *reliant chaque oiseau* [link each bird], asking the student to *make* this implication themselves (i.e., an additional cognitive process).

Although there were other instances where this occurred in the two versions, the experts did not point out all of them. For example, in the English version, Item 13 (nDIF) repeats an expression from the stem ("may not be able to tell if . . .," "would not help them to tell if . . .") whereas the French version uses different expressions in the two parts of the problem (e.g., *ne pas être capable de s'apercevoir que* [not being able to see] vs. *ne les aiderait pas à savoir* [would not help them knowing]). That is, the French version first focuses on the "ability to see that the dog is sick," but then asks about "knowing that it is sick." The first statement involves an observation whereas the second one involves knowledge. The same difference in logical relations of content was noted between Item 30 (nDIF) and 34 (nDIF), the former using *complexion* to translate "skin type" whereas the later uses *type de peau* [type of skin] (Paul).

Logical Difference in Probabilities In Item 10 (nDIF) Sabina noted a difference between the English and French versions related to the need for different forms of statistical reasoning. The item states:

> ENGLISH: On their drive to the lake, the Malas see a cliff that has many cracks in it and has many small rocks at its base.
> **Which is the most common cause of this type of damage to cliffs?**
> FRENCH: En se rendant au lac, la famille Mala observe qu'il y a de nombreuses fissures dans une falaise et qu'il y a beaucoup de petites roches au pied de celle-ci. [Going to the lake, the family Mala observes that there are numerous cracks in a cliff and that there are a lot of small rocks at the foot of it.]
> **Quelle est la cause probable de l'état de cette falaise? [What is the probable cause of the state of this cliff?]**

Sabina – as Paul – saw a difference between the expressions "type of damage to the cliff" and *état de cette falaise*. She first pointed out that "damage" refers to the fact that something happens to the cliff, whereas the French version uses *état* [state], and she pointed out that

damage and state (*état*) are not the same. She then detected a difference between "the most common cause" and *la cause probable* [the probable cause].

> Which is the most common cause, *Quelle est la cause principale de ce type* [which is the principal cause of this type], most common cause, *la cause probable, principale* [the probable cause, principal. *Quelle est la cause* [what is the cause]. *La cause la plus* [The cause the most] *La traduction, la cause probable, c'est pas tout à fait pareil que* « the most common ». *De l'état de la falaise* [This state of the cliff]. Of this type of damage to cliffs. *C'est comme si c'était une, quelque chose qui arrive, de ce type de dommage, donc ce n'est pas de l'état de la falaise* [It's like there was a, something that happens to, this type of damage, so it's not the state of the cliff]. (Sabina, 10, nDIF)

Although Sabina noted a difference, she did not quite put her finger on the structural difference, which is related to the fact that the English version refers to cliffs in general, whereas the French version specifically pertains to the cliff described in the stem (*cette falaise* [this cliff]). In semiotics, two very different types of reasoning are required in solving a problem (Eco, 1984). In deductive reasoning, the movement is from the general rule to the particular case, whereas in abductive reasoning, the movement is from a particular case to the (unknown) general rule. From a cognitive perspective on statistics, the English version asks for a comparison of the distribution of causes of a particular type of damage (population of cases); the French asks for the evaluation of the probable cause (*cause probable*) of this cliff (case). The two are different, as the most frequent cause in the population of such damage (cracks, lots of small rocks) does not necessarily mean the most likely cause for *this* cliff. (The most common murder weapon may not be the most probable in any specific case.) In more technical terms, the English version asks for the cause that leads to the highest number of cases of this type (p_{max}(data|cause)), whereas the French version asks for the cause that maximizes p given the data (p_{max}(cause|data)).

Cognitive–Conceptual Processes

In some instances, the two versions used different concepts. Thus, all three experts identified a difference between the English expression "size of the windows" and the French *la superficie des fenêtres* [surface of the windows]," but draw different conclusions as to the performance differences.

> *Parce que* size of the windows, *on parle pas de la superficie, on parle de la "taille," et ici on parle de superficie* [Because "size of window," we are not talking about *superficie* [surface area], we are talking about "size"]. (Sabina, 3, nDIF)
>
> So the difference in the (inaudible) the concept of size, the concept of *superficie*, which in English is area, right. So size and area is a bit . . . and especially if I have a window like that, right, or a window like that right, ok, so if I think area in my mind it's going to be more . . . than size, anyway. (Paul, 3, nDIF)
>
> I have noted is the difference between the word "size" in English which is fairly straightforward and understood by a most age groups and most students most people whereas *superficie en français* [surface (area) in French] might be more of more difficult for some students *la grandeur* probably would have been or might have been a better term in my mind. (Walter, 3, nDIF)

Sabina indicated not knowing whether this would make "a big difference," because length and width implied size, but suggested to replace the French term by *taille* [size]. Paul explicitly stated the difference between size, on the one hand, and *superficie* [surface (area)], on the other hand. He suggested using the term "area" in the English to produce an equivalent version of the item. Walter, noting the difference, did not articulate the conceptual difference, but commented on the difference in understanding that it implies for the two versions.

Conceptual features might be – and in some instances are – associated with DIF. For example, one item asks in the English version "What is the value of resistance to heat flow?" but in French *Quelle est la valeur de sa résistance à la perte de chaleur?* [What is the value of its resistance to the loss of heat?]. Here, the French version indicates a loss of heat, whereas the English version merely states that there is heat flow without indicating in the directionality of this flow. The same item states in English that the family "would like to insulate their basement," whereas the French version is about improving the state of the insulation (i.e., *améliorer l'isolation du sous-sol* [improve the insulation of the basement]). In the first instance, a specific case is described whereas in the second instance a state is referred to.

None of the experts anticipated performance differences on Item 40 (DIF_F), though Sabina noted a clear difference in the wording of the stem. In English, the question is "Why is a magnifying glass a poor tool to use when looking for sugar in a blood sample?," whereas the French question is about "not being a good instrument to research the blood [*pour rechercher le*

sang].” We thus find a conceptual difference between the two languages in the way the question is phrased.

Diversity Issues

The experts noted that the French version would particularly disadvantage Francophone students in minority settings (“I think there’s a bigger challenge for a Francophone in a minority setting than for English kids that’s for sure,” Paul), where the students tend to speak English not only with their friends after school, but, despite penalties, also at school, especially in unsupervised situations (Roth, 2005b). Paul noted that although they may speak French at home, these students are disadvantaged when compared with their peers in Quebec. He explained that these students are completely immersed in an Anglophone culture, so that some of their basic mechanisms will be in the context of English rather than French. He used an example from his own experience, where he tends to write quick notes to his wife in English rather than in the French language that they normally use. This could put tremendous pressure on such students in examination and testing sessions: “I’m a Francophone in my setting, my parents were English, Anglophone, I speak English all of the time and I see a word in an exam setting right, national exam, P.S.A., PISA exam, I’m nervous, and I see a word and another word so I panic” (Paul, 30, nDIF).

There were two instances where the experts noted diversity differences. Walter talked about the cultural diversity in his daughter’s class, where he would “find a very very rich multicultural mosaic where there “would be a Yessin, Mohammed, Nakhla, a Yasmin.” He referred to his own teaching experience as evidence that the cultural mosaic is richer in Francophone settings than in “predominantly Anglophone populations,” where “multicultural aspects is not as wide as in the Francophone program” and where it is “not uncommon to find large population from North Africa and the Middle East (inaudible) or other African part of the world.”

Another case where a diversity issue was noted occurred in Item 40 (DIF_F), where “we’re talking about the female teacher now we’re talking about a male teacher” (Paul). Not noted was

another such difference in Item 39 (DIF_F), where the English version presents Joe and Jill, but the French version contains two girls, Joanne and Jocelyn.

DISCUSSION

The outcomes of international and national cross-language studies of science achievement are used in policymaking and political-economical decisions in science education. However, there is evidence that such comparisons are not fair given that these tests are biased against one or the other linguistic group. This study was designed to investigate possible sources of DIF in English and French versions of a national achievement test. We drew on three experts, whose daily work included a large amount of verification of French translations of curriculum and examinations that certified translators had produced. Our descriptive quantitative analyses show that the experts denoted about an equal proportion of DIF and nDIF items (~70%) as exhibiting no or minimal differences. However, the protocol analyses provide deep insights into differences that operate along different dimensions and with different intensities, including (a) length of French versus English versions, (b) syntactic (grammatical) and semantic differences, (c) different logical structures in content or form of item, (d) differences in cognitive-conceptual content, and (e) diversity issues, which confirm the differences identified between multiple language versions of assessments in previous research.

As anticipated, there appeared to be a tendency to highlight different aspects associated with the expert's particular background and area of expertise. Thus, the two experts who had taught science – and therefore were able to draw on (content) pedagogical knowledge – pointed out aspects that would affect students in their understanding, the length of time to read (process information), fatigue, and, as a result, the effect this could have on performance levels. The expert who articulated more grammatical issues and fine details in the two languages did not highlight pedagogy related issues, in which she did not have background or experience. That is, the experts point out many important differences, and often agree on the particular issue raised as making the two versions different. What they differ on concerns the impact that might have on student performance and the degree of impact between the two versions. Thus, it is important

during test development or validation stages to include translation experts with a variety of backgrounds and experiences (i.e. those who have [content] pedagogical knowledge as well as experience working directly with students may notice different aspects of items that could potentially lead to DIF than those who have pedagogical knowledge but no direct experience with students).

These experts had been asked to participate because of their cultural competencies that English-native translators between the two languages often do not bring to the task. Thus, particularly in Francophone Quebec – the origin of Paul – there is a particular sensitivity with respect to the dominance of the Anglophone cultural and where Law 101 (Charter of French Language) has been instituted to preserve Francophone language and culture. However, of all the potential sources for DIF, diversity issues were the least frequent. This may indicate that the item designers already attended to the potential for cultural bias.

Our study has greatest family resemblance with another one that compared the equivalence of translations in international reading literacy studies (Arffman, 2010). That study noted (a) problems related to language-specific differences in grammar (e.g., word length, clause structure, reference and pronoun systems), (b) language-specific differences in writing systems (e.g., semantic use of commas), (c) language-specific differences in meaning (e.g., frequency of use of technical words in vernacular), (d) differences in culture (narratives situated in specific cultural settings, e.g., U.S. Deep South), (e) strategies used and choices made by the translators, and (f) writing errors (e.g., punctuation) (Arffman, 2010). In that study, the author conducted the analyses based on her theoretically informed interpretations of literary texts. Our study differed in that it drew on experts whose daily work involves making pragmatic choices of translations that they *knew* could significantly affect educational practice and assessment outcomes in their province. Moreover, the problems that plague literary texts, as indicated by two of the experts in this study (Paul, Sabina), are not the same as those that would operate in the science and mathematics texts that they tended to be responsible for. Also, the text of the individual test items tended to be very brief compared to the 196, 383, and 1,727-word texts that Arffman

analyzed. Our study shows that in addition to the six types of linguistic differences, which fall into our first, second, and fifth protocol-derived theme, our experts also noted differences in the logical structure within the content and the relation between items, as well as differences in cognitive-conceptual content. Thus, by examining the think-aloud protocols of translation experts as they assessed the equivalence of translated items this study has confirmed and added to the existing understanding of the types of differences that are found in translated items that may account for differential functioning across language groups. Only empirical studies that use think-aloud protocols for investigating thought *processes* while students take tests with items of known DIF and non-DIF status – one of which we are in the process of setting up and conducting at the time of this writing (Summer 2012) – can she further light on the question whether what the expert identified as possible sources are *actual* sources of variation in thinking about and responding to test items. Further investigations comparing other language combinations should be designed to better understand possible linguistic sources of DIF in multi-lingual science assessment.

Further, this study shows that even when the content of text is technical and conceptual, there are difficulties in making translations equivalent. For example, the experts pointed out the interrelation between expression and conceptual content in numerous cases. For example, the French term *êtres vivants* necessarily implies beings, animals and humans, which is not equivalent to the English "living *things*," which include plants. Similarly, the French verb *foudroyer* applies to humans but not to rocks and cliffs, whereas the English "lightening strikes" applies to both. We cannot, therefore, get at some abstract common core that would make items *exactly* equivalent in their conceptual science content as stated in two languages. Any attempt to overcome this problem, therefore, is doomed to fail. This is so because the problems in the construction of equivalent translations of science test items is endemic to *any* translation produced between the two versions (Derrida, 1996; Ricœur, 2004). In fact, these authors state that even within-language translations never are equivalent. This is apparent in the common observation that a student may understand a question when asked in one way but not when the

"same" question is asked in another way (e.g., Roth and Radford, 2011). This point of non-equivalence *within* a language also is visible in the attempts of our experts to produce alternative translations to simplify a French expression or construction. That is, two versions of French may be equivalent in "meaning," but they would not be equivalent in terms of many other dimensions, including length, time it takes to process, grammatical structure, cognitive difficulty, and so on. This implies that the results of tests such as PISA should be taken with a grain of salt when it comes to the signification of differences between countries and to the use of such differences in policymaking, curriculum design, and science teacher training.

The experts in this study, working individually, did not consistently differentiate between items that exhibit DIF and those that do not. Future research may be designed to specifically investigate (a) whether groups of experts with differential levels of expertise in translation, pedagogy, and cognitive psychology identify DIF items to a greater degree than our experts working alone and (b) the degree to which those differences that experts (expert groups) identify play a role in the actual processing of the items in the two languages. In the first situation, the experts would inherently make their thinking available for each other, decreasing the level of artificiality that experts might associate with the think-aloud protocol (they would not have to be encouraged to read and think aloud). In our research group, one such study is currently in the planning stage. In the second type of research, the think-aloud protocol might be used in conjunction with pair- or group-wise administration of items, which, again, would inherently encourage participants to make available their reasons to their collaborators.

In summary, this study explored the think-aloud protocols of expert translators while they engaged in the act of evaluating the equivalence of translated versions of a multi-lingual large-scale assessment. Through this study we have gained insight into the characteristics of the items the translators attended to and thought were critical to equivalence. Five major dimensions were revealed as potential sources of differential functioning. By paying particular attention to these dimensions during initial translation and verification stages of test development, test developers

may be able to identify and remove potential sources of differential functioning prior to the administration of the test.

REFERENCES

Arffman, I. 2010. Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research, 54*: 37–59.

Artelt, C. and Baumert, J. 2004. Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs [On the comparability of student performance in reading tasks of different language origins]. *Zeitschrift für Pädagogische Psychologie, 18*: 171–185.

Benjamin, W. 1972. Die Aufgabe des Übersetzers [The task of the translator], in *Gesammelte Schriften Bd. IV*. Frankfurt/M, Germany: Suhrkamp-Verlag.

Bond, L., Moss, P. and Carr, P. 1996. Fairness in large-scale performance assessment. In *Technical issues in large-scale performance assessment*, Edited by: G. W. Phillips, 117–140. Washington, DC: National Center for Education Statistics.

Bybee, R., Fensham, P. J. and Laurie, R. 2009. Scientific literacy and contexts in PISA 2006 science. *Journal of Research in Science Teaching, 46*: 862–864.

DeBoer, G. E. 2011. The globalization of science education. *Journal of Research in Science Teaching, 48*: 567–591.

Derrida, J. 1996. *Le monolinguisme de l'autre ou la prothèse d'origine* [Monolingualism of the Other; or, The prosthesis of origin]. Paris, France: Galilée.

Eco, U. 1984. *Semiotics and the philosophy of language*. Bloomington, IN: Indiana University Press.

Ercikan, K. 2002. Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing, 2*: 199–215.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G. and Koh, K. 2004. Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*: 301–321.

Ercikan, K., Arim, R.,G., Law, D. M., Lacroix, S., Gagnon, F. and Domene, J. F. 2010. Application of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice, 29*: 24–35.

Ercikan, K. and Roth, W.-M. 2006. What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35* (5): 14–23.

Ericsson, K. A. and Simon, H. A. 1993. *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Geisinger, K. F. 1994. Cross-cultural normative-assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*: 304–312.

Grisay, A. and Monseur, C. 2007. Measuring the equivalence of item difficulty in the various versions of an international test. Studies in Educational Evaluation, 33, 69–86.

Hambleton, R. K., Merenda, P. F. and Spielberger, C. D. 2005. Eds. *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H. and Rogers, J. 1991. *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Krippendorff, K. 2012. *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.

Lee, J. F. 1986. On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8*: 201–212.

Olsen, R. V., Prenzel, M. and Martin, R. 2011. Interest in science: A many-faced picture painted by data from the OECD PISA study. *International Journal of Science Education, 33*: 1–6.

Rankin, J. M. 1988. Designing thinking-aloud studies in ESL reading. *Reading in a Foreign Language, 4*: 119–132.

Rey, A. 2012. *Le grand Robert de la langue française* (version électronique [The grand Robert of the French language, electronic version]. Accessed February 21, 2012 at http://gr.bvdep.com/

Ricœur, P. 2004. *Sur la traduction* [On translation]. Paris: Bayard.

Roth, W.-M. 2005a. *Doing qualitative research: Praxis of method*. Rotterdam, The Netherlands: Sense Publishers.

Roth, W.-M. 2005b. Telling in purposeful activity and the emergence of scientific language. In *Establishing scientific classroom discourse communities: Multiple voices of research on teaching and learning,* Edited by: R. Yerrick & W.-M. Roth, 45–71. Mahwah, NJ: Lawrence Erlbaum Associates.

Roth, W.-M. 2009. Limits to general expertise: A study of in- and out-of-field graph interpretation. In *Cognitive psychology research developments*, Edited by: S. P. Weingarten & H. O. Penat 1–38. Hauppauge, NY: Nova Science.

Roth, W.-M. and Hsu, P.-L. 2010. *Analyzing communication: Praxis of method*. Rotterdam, The Netherlands: Sense Publishers.

Roth, W.-M. and Radford, L. 2011. *A cultural-historical perspective on mathematics teaching and learning*. Rotterdam: Sense Publishers.

Schoenfeld, A. 1985. *Mathematical problem solving*. Orlando, FL: Academic Press.

Table 1: Summary statistics of the DIF TAPs

	Paul				Sabina				Walter			
	Identified Differences				Identified Differences				Identified Differences			
	Yes		No/Min		Yes		No/Min		Yes		No/Min	
	Performance D		Performance D		Performance D		Performance D		Performance D		Performance D	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
DIF*	5	0	1	14	0	1	0	19	2	3	0	15
Non-DIF	2	1	3	14	0	1	0	19	3	3	0	14

*DIF items that were previously identified as having linguistic differences

Table 2: Levels of Agreement as to the Nature of Items

	DIF		non-DIF	
No. Exp	Y	N	Y	N
3	0	10	1	11
2	4	6	0	8

Table 3: Number of DIF Items Identified as DIF Item

	Number of Experts Identifying the DIF Item as DIF Item	
Type of DIF item \	2	1
DIF_F (n = 9)	0	5 (2)[1]
DIF_E (n = 11)	3 (2)	1 (1)

[1] In parenthesis are stated the number of the items labeled to be clearly different that are expected to lead to performance differences.

Table 4: Summary of categories identified in the think-aloud protocols with examples

Category	**Explanation** (Expert refers to . . .)	**Example**	**Code**
Length/Time	the amount of time it takes to do the item or to the length of the text	*Nos traductions sont toujours plus longues en français qu'en anglais* [Our translations are always longer in French than in English] (Sabina, 17, nDIF)	1
Linguistic Differences			2
General Reading	overall reading level of the text	That's a really nice literary French structure, okay. You never use that in day to day French. (Paul, 36, nDIF)	2a
Unfamiliar	a term or expression being unfamiliar to the students	*La falaise a été foudroyée par un éclair* [The cliff has been struck by a lightning bolt], how many student would be familiar with *foudroyée* [struck]? (Walter, 10, nDIF)	2b
Grammar	more complicated sentence structure or different tenses	*Le temps est diffèrent. Ici* (French) *on parle du futur et le présent, et là* (English) *c'est le pass*é. [The tense is different here [French] one speaks of the future and the present, and here (English) it is the past. (Sabina, 34, nDIF)	2c
Figure/Ground	differences in the way actions and context are related	*À la maison* in French mean "at home" right? "Around the house" is a little different. Okay say, "working around the house" and all that . . . this feels like home *à la maison* means home, right. (Paul, 1, DIF_F)	2d
Logical Structure			3
Negations	to negations, double negatives, and	when you use *la plus forte réduction,* reduction is this,	3a

	inversions	*la plus forte* is this right. The strongest reduction. (Paul, 4, DIF_E)	
Forms of Inclusion	concepts used in the two languages imply the same or different levels of inclusion	they use for skin color they use the word complexion, but complexion is English right, English word, skin complexion . . . may imply more than color. (Paul, 30, nDIF)	3b
Logical Relations	different relations between items or within items between objects	So *le mur du sous-sol est construit de quince centimetres d'épaisseur de briques . . . quince centimetres d'épaisseur de briques* . . . that's a very complex way to say fifteen-centimetre brick wall. (Paul, 20, DIF_E)	3c
Probabilities	different probabilities required	*La traduction, la cause probable, c'est pas tout à fait pareil que* "the most common." *De l'état de la falaise* [This state of the cliff]. (Sabina, 10, nDIF)	3d
Cognitive Conceptual	conceptual differences between two terms presented as equivalent (e.g., surface vs. size)	Parce que « size of the windows » on parle pas de la superficie on parle de la taille, et ici on parle de superficie. [Because "size of window," we are not speaking of surface, we are speaking about size, and here we are speaking about surface] (Sabina, 3, nDIF)	4
Diversity	differences in the culture or gender	[The test items] would benefit of little bit more from sort of a variation of first name that mirrors in particular the reality of Francophone minority where minority situations. (Walter, debriefing)	5