International Scholarly Research Network
ISRN Education
Volume 2012, Article ID 179824, 11 pages
doi:10.5402/2012/179824

Research Article

International Education Studies: Increasing Their Linguistic Comparability by Developing Judgmental Reviews

Inga Arffman

Finnish Institute for Educational Research, University of Jyväskylä, Jyväskylä, P.O. Box 35, 40014, Finland

Correspondence should be addressed to Inga Arffman, inga.arffman@jyu.fi

Received 13 December 2011; Accepted 3 January 2012

Academic Editors: B. Marlow and P. A. Prelock

In international education studies, the different-language test versions need to be equally difficult to read and answer for the test to be valid. To ensure comparability, several quality control procedures have been developed. Among these, surprisingly little attention has been paid to judgmental reviews and their ability to identify language-related sources of bias. Also, the reviews have often failed in identifying biases. This paper explored whether it is possible to improve the ability of judgmental reviews to identify language-related sources of bias. A new review was made of two Finnish items which in the PISA (Programme for International Student Assessment) 2000 reading test showed differential item functioning but for which no clear language-related explanations were found in the review in 2000. The items were compared systematically, at all linguistic levels, to the corresponding items in the English and French source versions, at the same time taking into account the cognitive processes required to answer them and students' written responses to them. Language-related explanations were found for both items which may have led to differences in performance, suggesting that it is possible to make judgmental reviews better able to identify language-related bias. Suggestions are given on how to do this.

1. Introduction

International education studies are conducted on a regular basis to assess knowledge and skills across countries. In these studies, a common test is usually translated into the languages of all participating countries. When this is the case, it is important to ensure that the different-language test versions are comparable, or equivalent, to each other: that they measure the same construct at a comparable level of difficulty. This requires that the items in the versions are not biased, unfairly favoring some countries or language groups over others, thereby making it easier for some respondents to find the correct answer. A failure to establish equivalence jeopardizes the validity of inferences made on the basis of the results of the test.

For the items to be equivalent, the mental effort required of testees to respond to them needs to remain the same across languages. No version must place a heavier cognitive load ([1]; see also e.g., [2, pages 93-4]), or consume more of the limited processing capacity of testees' working memory compared to other versions. This requires, among other things, that all different-language items and stimulus texts be equally easy to read and understand. If this is not the case and some items or stimuli are harder to understand than others, more working memory is needed to decode and make meaning of them. Importantly, when reading and comprehension are the first phase in the survey response process, there is less memory available for actually responding to the items (i.e., retrieving the relevant information, forming the judgment, and editing the answer; e.g., [3, page 8]). Readers of those versions requiring greater reading comprehension prior to responding would then be at a disadvantage.

To ensure equivalence, several quality control procedures have been developed. An important part of these procedures are judgmental reviews conducted by expert reviewers after the test has been administered (i.e., *a posteriori*) to find the items that are biased. The reviews are preceded by statistical item analyses, which detect the items that show differential item functioning (DIF). An item shows DIF, when respondents of equal ability but from different groups do not have an equal probability of responding to the item correctly. From among the DIF items, the reviewers identify

those that they believe are biased. These items are then deleted from the analyses or corrected for use in future studies.

In international education studies, an important potential source of DIF and bias are language-related differences (other sources including, e.g., cultural and curricular differences, often overlapping with language-related differences; for cultural and curricular differences, see e.g., [4–6]). Language-related differences refer to differences between languages and translation errors and can be roughly classified as in Table 1 (cf. [7–13]). Of the six language-related sources of bias, the first four typically derive from differences between languages whereas the last two are concerned with faulty or inappropriate translation.

Judgmental reviews thus play a significant role in ensuring linguistic comparability and validity. Compared to other quality control procedures, however, relatively little attention has been paid to these reviews. For example, often no clear standard or instructions on how to conduct the reviews exist. Research also shows that often the reviews have been unable to identify sources of DIF consistently and accurately. For example, it has been detected that reviewers have frequently been unable to correctly predict DIF and the direction of DIF (e.g., [14–16]). Ercikan et al. [11] found that reviewers' judgments as to which items are biased and why are frequently incorrect and, therefore, need to be verified by checking them against what the testees themselves think.

The above also seems to apply to international education studies and to the identification of language-related bias. For example, concerning the reviews conducted in the PISA 2000 test, it was reported that the International Project Centre, responsible for making the final decisions on whether to delete or retain items, had to base its decisions on highly deficient data, because the information provided by participating countries on their reviews was often either missing altogether or was extremely vague, with no clear explanations for the poor functioning of the items (A. Grisay, personal communication, February, 24, 2009).

Moreover, when examining the amount of DIF within and between language groups in the PISA reading test, Grisay and Monseur [17] found that the amount of DIF varied considerably depending on the proximity of the language. When comparing versions sharing the same language (e.g., all English versions), the proportion of DIF items was very low (3% on average). When comparing versions using different Indo-European languages (e.g., English versus French versions), it was considerably higher (25–30%). However, the highest proportions of DIF items (ca. 45%) were found when comparing the relative item difficulties between, for instance, the English and French versions, on one hand, and the most distant, non-Indo-European (Asian) versions, on the other.

Similarly, Grisay et al. [18] examined communalities of item difficulties and the mean absolute magnitude of DIF between national and international item parameters in the PISA 2000 and the PIRLS 2001 study. In both studies, a significant item-by-language interaction was discovered, in that in both studies items were found that behaved in a deviant way in all or most translations into a same language (e.g., all French versions). Also, the mean absolute magnitude of DIF was higher in instruments translated into more distant (Asian) languages than in instruments translated into Indo-European languages. The findings suggest that there were language-related differences in the versions (although cultural and curricular explanations could not be entirely excluded) but that these went unnoticed and/or were not identified as sources of bias in the judgmental reviews (see also [8, 19]).

That there were language-related differences in PISA and PIRLS versions which, however, were not identified as sources of bias in judgmental reviews seems to be supported by the extremely small number of items identified as biased and deleted in the studies. For example, in the PISA reading test, a total of 141 items were administered. Of these, 7 were deleted in all countries, because they showed poor psychometric characteristics in more than 8 countries ([20, pages 262–5]; see also [21, page 101]). The reasons for the deletions included not only language-related differences, but also, for instance, cultural differences and printing errors (e.g., [12, 22]). In addition, 21 items were deleted at the national level [23, page 153], the reasons again ranging from translation issues to printing errors (e.g., [12]). National deletions were made in 17 countries, with the vast majority of these countries showing only one deletion (Austria; Canada, Fr; Germany; Greece; Hungary; Iceland; Mexico; Poland; Russian Federation; Spain; Sweden) or two deletions (Belgium, Fl; England; Italy; The Netherlands); in German-speaking Switzerland 5 and in Korea 4 deletions were made. However, in most countries there were no national deletions. This was the case, for example, in all Asian countries, except Korea, and in Finland. [23, page 153]. Thus, in the vast majority of PISA countries, including those in Asia (and Finland), extremely few items were identified as showing language-related bias, although other research suggests that there were such items in these versions.

2. Purpose and Scope of the Study

This study examined whether it would be possible to improve the ability of judgmental reviews conducted in international education studies to accurately and reliably identify language-related sources of bias (testing whether or to what extent the sources of bias actually led to differences in performance and bias was beyond the scope of this study). The purpose was to help to develop judgmental reviews.

The study focused on biases caused by language-related differences, because they are the most obvious and most common type of difference and because they also seem to have been most problematic differences to identify and/or to deal with. However, slightly modified, the problems discussed and suggestions provided also apply, for example, to other sources of bias. Also, as an example of judgmental reviews in international education studies, the study used the review conducted in Finland in the PISA 2000 reading test (because on it most data were available). The study thus focused on reading literacy studies and the skills, strategies, and processes assessed in these studies. However, since the way in which studies in other subjects and their quality control procedures (e.g., translation and adaptation) are

Table 1: Language-related sources of bias.

(1) Differences in grammatical form and structure (e.g., word order, complexity of syntax, and pronouns), including typographical differences (e.g., punctuation, capitalization).
(2) Differences in meaning, connotations, and register, especially as concerns key words (e.g., metaphors, idioms; words with several meanings in one language but not in the other; technical words translated as more everyday words).
(3) Differences in item format (e.g., sentence-completion items).
(4) Words and structures prone to misreading (e.g., the French *au-dessous* "below" read as *au-dessus* "above").
(5) Inaccurate translation, including additions or omissions (e.g., omitting the phrase "according to the text" from the item, in which case responses may be given not on the basis of the text but on the basis of general knowledge) and differences in how literal the match between the item and the stimulus text is (e.g., literal versus synonymous match).
(6) Translation errors (e.g., semantic, syntactic, and orthographic errors).

implemented is usually more or less the same as in reading tests (see e.g., [24, 25]), the study may be expected to apply largely also to these other studies.

The paper first briefly describes PISA and its quality control procedures, paying special attention to judgmental reviews, then depicts how the present study was conducted, and, finally, summarizes the findings, providing suggestions on how to develop judgmental reviews and where to focus in future research.

3. Quality Control Procedures in PISA

PISA is a regular, ongoing assessment administered to 15 year olds. The studies are conducted every three years, each study being preceded by a field trial. During every assessment cycle, the analyses focus on one of the three major domains of PISA: reading literacy, mathematical literacy, or scientific literacy. In PISA 2000, the major domain was reading literacy. PISA 2000 was the first PISA reading literacy study, and it was conducted in 43 countries.

PISA reading tests consist of units made up of one or more stimulus texts and two or more question items. The stimulus texts are authentic texts, representing different text types, genres, forms, lengths, and degrees of difficulty. The items are either multiple-choice or constructed response items and focus on three reading strategies: retrieving information, interpreting texts, and reflecting on and evaluating the content or form of a text [26].

In the following, a brief overview is provided of some of the quality control procedures followed in PISA to ensure equivalence.

3.1. Procedures Preceding Judgmental Reviews

3.1.1. Development of the Source Versions. Before the field trial, each participating country can propose texts and items to be included in the source instrument. These are exposed to extensive reviews by test developers, subject experts and translators, and to prepilots in some participating countries. Each country also reviews the items from the point of view of cultural relevance, interest, sensibility, and translatability. In more recent PISA studies, moreover, some of the items have been subjected to cognitive interviews (interviews with testees to obtain data on their response processes, e.g., how they understand items and why they answer as they do) and think aloud protocols (or TAPs, where subjects think aloud while performing a task). Finally, the test developers decide on the English and French source versions to be used in the field trial [27].

3.1.2. Translation of the Target Versions. Each participating country makes its national versions on the basis of the English and French source versions. The recommended procedure is double translation from two parallel source texts, followed by national reconciliation and international verification: based on the two source versions, two national translators produce two independent versions in the target language. These two versions are reconciled by a third translator into one national version, which is verified by a fourth translator, appointed and trained by the International Project Centre [27]. (The main difference between the PISA translation procedure and the procedure followed in PIAAC is that in the latter only one (English) source version is produced. The IEA procedure, for its part, differs from the PISA procedure in that the test is provided only in English and that only one translation is produced from every source version.)

In Finland, the translation procedure has deviated somewhat from what has been recommended, in that in the first phase only one translation has been made. The translations have been made from English and then revised and reworked by other translators during the following phases. In PISA 2000, the French version was only employed for some minor cross-checks, whereas in PISA 2009 the cross-checks were systematic.

Translators are provided with detailed translation guidelines on how to avoid translation traps that often lead to bias (for a fuller account of the guidelines, see [28]). They are also given more item-specific translation notes, helping them with more specific translation or adaptation problems. For example, notes may be used to ask the translator to imitate, if possible, the stylistic characteristics (e.g., irony) of the source text. Finally, the translators receive information on the reading strategies required to answer the questions (i.e., retrieval, interpretation, reflection, and evaluation). This helps them to make sure that the nature of the item and its level of difficulty remain intact. For instance, when told that an item focuses on interpretation, the translator knows that special attention needs to be paid to keeping the level of explicitness unchanged.

3.1.3. Statistical Item Analyses. After the field trial, the International Project Centre conducts statistical analyses to identify the items that show DIF. The analyses are performed using a mixed coefficients multinominal logit model, which is a Rasch-based form of item response modeling. The results of the analyses are compiled into national reports containing all "dodgy" items, or items that are flagged for that country for one of the following reasons: their difficulty is significantly lower or higher than average; one of the nonkey categories in a multiple-choice item has a positive point-biserial correlation, the key category point-biserial correlation is negative, or the category abilities for partial credit items are not ordered; or the discrimination of an item is small or large (for a fuller account of the statistical analyses and the reports, see [21]).

3.2. Judgmental Reviews. After the statistical analyses, the reports on the flagged items are sent to the participating countries, which are to conduct judgmental reviews on them and to identify the items that have, for example, translation or printing problems. However, no clear directions are provided as to what is meant by translation problems, who should conduct the review, and how, for example (e.g., [21, pages 99–108]; [29]).

Some directions, nevertheless, are given in the manuals provided for national project managers (NPMs) before each study. For instance, the PISA 2000 (Main Study) Manual [30, pages 37–9] exemplified translation (and printing) problems by listing four types of such problems: omissions or additions of information; faulty matches between an item and the stimulus text (e.g., a literal match in the target version but a synonym in the source versions); differences between languages, because of which nuances of meaning are difficult to transfer across languages; flaws in graphic presentations. NPMs were advised to look for problems such as the above in their national versions and to correct them before the Main Study. In later NPM manuals (e.g., [31]), NPMs have been encouraged to correct more or less exclusively only outright translations errors; no mention has been made of differences between languages, for example.

In the PISA 2000 (Main Study) Manual [30], no instructions were provided on who should conduct the review. In the PISA 2000 Field Trial manual [28, pages 28-9], NPMs were encouraged to conduct the review together with translators. In later PISA Manuals (e.g., in PISA 2009), however, the recommendation has been to have the items analyzed by domain experts familiar with testing (e.g., [31]). In Finland, the PISA 2000 judgmental review was conducted by a reading literacy researcher who was also a mother-tongue (Finnish) teacher. The reviewer had headed the scoring of Finnish students' open-ended questions and therefore knew how they responded to the questions.

After the countries have analyzed the items, the results are communicated to the International Project Centre. In PISA 2000, no directions were given on what and how to report. In Finland, however, a free-form report was written which contained all the flagged items and possible explanations for their poor functioning (Finnish Institute for Educational Research (FIER), [32]). The explanations varied slightly in degree. However, they were all relatively cautious, containing expressions such as "The problem with this item was probably…" and "Perhaps this could at least partly explain…" and in one case even extremely vague, with no clear indication as to whether the problem was the translation or the item itself. The directions on how to report have remained more or less the same also in all later PISA cycles. However, a simple electronic form, or box, has been developed by means of which the results have been communicated to the Project centre (e.g., [31]). Finally, on the basis of the national results, the Centre decides which items are deleted from the final analyses (either internationally or nationally), which items will be used in the Main Study, and whether revisions still need to be made to them. At this stage, no cognitive interviews are conducted.

Similar statistical analyses and judgmental reviews are also conducted after the Main Study. On the basis of these analyses, the Project Centre decides which items have to be removed from the final analyses or corrected for future use.

4. Research Design

This study is a substudy of a larger action research study, aimed at developing procedures promoting translation equivalence in international education studies. Action research is the process by which practitioners study their own problems scientifically in order to evaluate and improve their practice ([33], [34, page 5]). In action research, the researcher often takes an involved role as a participant in the process. The researcher in this study has acted as a reviewer in Finland in some later PISA studies (but was in no way involved in the PISA 2000 study).

4.1. Data. The principal data of this study consisted of two PISA 2000 reading items which in 2000 were identified as harder in Finland than expected (i.e., harder than the average of all countries) but not interpreted as being biased because of language-related differences. The items were thus retained in the final analyses.

There were a total of 52 flagged items in Finland, and of these, 40 displayed DIF and were harder or easier in Finland than expected [29]. In the judgmental review in 2000, six of these 40 DIF items were interpreted as potentially having language-related explanations [32], and of the six items, two were deleted internationally, because they showed poor psychometric characteristics in several countries. However, for the remaining four items, the International Project Centre did not find sufficient grounds for deletion.

For reasons of confidentiality, the two items actually analyzed in this study had to be selected from among the 11 released PISA reading units. In these 11 units, there were 48 items, and of these, 11 were identified as unusually hard or easy in Finland [29]. In four of these 11 items, the DIF was interpreted in the judgmental review in 2000 [32] as possibly due to a language-related difference, and of these four items, two were among the items that were deleted internationally [23]. In the released units, there were thus 9 items which in Finland were identified as DIF but which were not interpreted as being biased because of

Table 2: Method for conducting the Judgmental review.

(1) Carefully reading the items and the texts with which they occurred and getting acquainted with the processes and strategies assessed in the items.
(2) Consulting the report on Finland's dodgy items [29] and the national judgmental report 2000 [32].
(3) Comparing the Finnish items and texts to the corresponding items and texts in the English and French source versions. The comparisons were made systematically at all linguistic levels (morphology, syntax, semantics and pragmatics), paying special attention to the sources of translation-related DIF presented in Table 1. The linguistic elements were compared against the cognitive processes required to respond to the items and the factors that have been found to affect item and text difficulty, such as familiarity, explicitness and transparency (e.g., [8, pages 40–4] [36–39]). The objective was to explore whether there were linguistic elements in the Finnish versions which differed so much from the corresponding elements in the source versions that they could have changed the cognitive content and/or difficulty of the items.

language-related differences and which were not deleted from the final analysis.

A preliminary review [35, page 234] was conducted of these 9 items. On the basis of this review, two items were selected for further, in-depth discussion in this study. The two items were selected, because—in line with the principle of purposeful sampling—they were considered to provide a representative picture of the typical language-related differences found in the DIF items and the judgmental processes needed to analyze them (for a fuller description of the two items, see Section 5 and http://www.pisa.oecd.org/dataoecd/44/62/33692744.pdf).

4.2. Method. To examine whether it would be possible to improve the ability of judgmental reviews in international education studies to accurately and reliably identify language-related sources of bias, a new review was made of the two PISA 2000 items which in Finland showed DIF but which were not interpreted as being biased because of language-related differences. The review was conducted as described in Table 2.

Mainly the review consisted of systematic linguistic comparisons between the source and Finnish versions. However, to gain more reliable data on the reasons for the poor functioning of the items, two documents were consulted: the report on Finland's dodgy items [29], which showed how the items functioned in Finland and which alternatives Finnish students chose in multiple-choice items; even more significantly, the national review report 2000 [32], which contained information on what Finnish students actually said (i.e., wrote) in open-constructed items (cf. [40, page 195]). This information provided indirect data (or clues) concerning the students' cognitive processes (e.g., what factors may have affected their answering; cf. [11]), thereby making it possible to make more informed guesses about the sources of DIF. The latter document also revealed whether any language-related explanations were suggested by the reviewer for the poor functioning of the items.

The review was made by two independent reviewers: me, who is a linguist and researcher specializing in test translation and reading literacy and have acted as a reviewer in later PISA rounds, and a linguist and translator competent in Finnish (native), English, and French and with some experience in test translation. Our reviews differed in that only I conducted the review in its entirety (stages 1 to 3), whereas the other reviewer worked blinded, not knowing how the items functioned in Finland (e.g., whether they were unusually hard or easy; i.e., skipping stage 2). Our judgments were in all respects unanimous.

5. Review

5.1. Item 1. Item 1 was a multiple-choice item occurring with an expository magazine article on sport shoes originally published in a French-Belgian magazine for adolescents. In the item, students were presented with a sentence quoted verbatim from the stimulus text. The sentence was made up of two parts, and students were to infer what the relationship between these two parts was. There were, supposedly, no explicit markings, such as conjunctions, to signal the meaning relationship. The item is presented below in English, French, Finnish, and as back translated from Finnish into English (Table 3).

The correct alternative was D, that the second part provided the solution to the problem that was described in the first part. Internationally, this problem-solution relationship was easy to find. In Finland, however, the item proved harder than expected. In the FIER [32] it was mentioned that the item was a "difficult task … about relations between sentences," apparently meaning that the item was difficult because it was about relations between sentences. However, since no further explanations were provided, it is impossible to know whether the judgment meant that items about relations between sentences are always difficult (at least for Finnish students) or that these relations were especially difficult in this item and for Finnish students in particular. If the latter is the case, it seems to imply that there was something in the Finnish item which made it difficult when compared to the items in the source versions.

From the OECD national report [29] it can be seen that in Finland alternative C was more popular than in other countries, suggesting that many Finnish students had difficulty differentiating between alternatives C and D. A close comparison of the question stems and alternatives C and D between the English (with the French version basically similar) and Finnish versions shows that there were grammatical differences between the versions, because of which inferring the problem-solution relationship and selecting the correct answer may have been somewhat harder for Finnish readers.

The first difference concerns the question stem and its signaling of meaning relations and its word order. Even

Table 3: Item 1 in English, French, Finnish, and as back-translated from Finnish into English.

English source version	*French source version*
Look at this sentence from near the end of the article. It is presented here in two parts: "***To** avoid* minor but painful conditions such as blisters or even splits or athlete's foot (fungal infections)…" (first part) "…the shoe must allow evaporation of perspiration and must prevent outside dampness from getting in." (second part) What is the relationship between the first and second parts of the sentence?	Examinez la phrase suivante, qui figure vers la fin de l'article. Elle est présentée ci-dessous en deux parties: "***Pour** éviter* les ennuis de parcours mineurs, mais douloureux-cloques et ampoules, voire crevasses ou mycoses (champignons)…" (premiere partie) "…la chaussure doit permettre l'évaporation de la transpiration et empêcher l'humidité extérieure de pénétrer." (seconde partie) Quelle est la relation entre la première et la seconde partie de cette phrase?
The second part (a) contradicts the first part (b) repeats the first part (c) illustrates the *problem* described in the first part (d) gives the solution to the *problem* described in the first part	La seconde partie de la phrase: (a) contredit la première partie (b) répète la première partie (c) illustre le *problème* décrit dans la première partie (d) donne la solution au *problème* décrit dans la première partie
Finnish translation	*Back translation from Finnish*
Katso tätä artikkelin loppupuolelta otettua virkettä. Se esitetään tässä kahdessa osassa: "Vähäisten mutta kivuliaiden vaivojen, kuten rakkojen tai jopa haavaumien tai "urheilijan jalan" (sienitulehduksien) *välttämise**ksi**…*" (ensimmäinen osa) "…kengän täytyy sallia hien haihtuminen ja estää ulkopuolista kosteutta pääsemästä sisään." (toinen osa) Mikä on virkkeen ensimmäisen ja toisen osan suhde?	*Look at this sentence taken from near the end of the article. It is presented here in two parts: "Minor but painful conditions, such as blisters or even splits or "athlete's foot" (fungal infections) *avoid* + ***translative suffix**…*" (first part) "…the shoe must allow evaporation of perspiration and prevent outside dampness from getting in." (second part) What is the relationship between the first and second parts of the sentence?
Toinen osa (a) on ristiriidassa ensimmäisen osan kanssa (b) toistaa ensimmäisen osan (c) havainnollistaa ensimmäisessä osassa kuvattua *ongelmaa* (d) esittää ratkaisun ensimmäisessä osassa kuvattuun *ongelmaan*	The second part (a) is in conflict with the first part (b) repeats the first part (c) illustrates the in the first part described *problem* (d) gives the solution to the in the first part described *problem*

* shows that what follows is somehow incorrect (e.g. grammatically).

though there are in the source text sentence no conjunctions to explicitly signal the problem-solution relationship, the sentence does contain an explicit marker referring to another meaning relationship. The marker is the preposition *to* (in French *pour*), which, however, is here used conjunctively, and the relation which it signals is that of purpose [41, pages 236-7]. The relationship of purpose, in turn, is logicosemantically close to the problem-solution relationship in that both imply a causal rather than, for example, a replacive, appositive, or exemplificatory relationship [41, pages 225–41], the relationships suggested by alternatives A, B, and C, respectively. The relationship of purpose, moreover, appears as the very first element in the long sentence.

However, in the Finnish sentence there are no prepositions, conjunctions, or any other words to refer to purpose or any other causal relationship. Instead, the relation of purpose is denoted by the suffix *-ksi-*, which, in addition, is agglutinated to the end of its head (*välttämise**ksi*** "avoid + translative"). This, again, is because whereas in more analytic languages, such as English and French, meaning relations are generally indicated by free morphemes, such as prepositions [42], in more synthetic languages, such as Finnish [43], they are mostly conveyed by bound morphemes, such as case endings affixed to word stems. In addition, in Finnish, nonfinite clauses of purpose are not made up of verbs, as in English and French, but of nouns. When these nouns have modifiers, the modifiers are, in line with the Finnish preference for premodification [44, pages 242–9], placed before the noun. These sometimes result in heavily left-branching structures. This is also the case in Item 1, where *välttämiseksi* "to avoid" is preceded by 13 words.

Thus, in both source versions, a separate word is used in the question stem to indicate the relationship of purpose, and the word, moreover, occurs sentence-initially. However, in the Finnish version, the relationship is denoted by a less explicit suffix which is affixed to its head word and which occurs as the final element in a long and heavily left-branching noun phrase—and which therefore easily gets lost amid all the other, less important information. In the Finnish version, the causal relationship thus does not appear as transparent and as easily inferable as in the source versions, which, in turn, may have made finding the correct alternative harder for Finnish readers (especially those with less developed reading skills). Also, because of the heavy premodification, an extra burden seems to have been placed on the memory capacity of Finnish readers, who had to keep

Table 4: Item 2 in English, French, Finnish, and as back translated from Finnish into English.

English source version	*French source version*
Here are some of the early references to the panther in the story. "the *cry* awoke her, a sound so anguished…" (line 32) "The answer was a repeated *cry,* but less shrill, tired sounding…" (line 44) "She had…heard their *cries,* like suffering, in the distance." (lines 52-53) Considering what happens in the rest of the story, why do you think the writer chooses to introduce the panther with these descriptions?	Voici quelques-unes des premières références au puma dans le récit: "le *hurlement* la réveilla dans la nuit, un cri si angoissé … ": (ligne 40) "La réponse fut un autre *hurlement,* moins perçant celui-là, comme fatigué, … " (ligne 53) "elle les avait entendus pousser leurs *cris* au loin, comme des cris de souffrance." (lignes 62-63) En tenant compte de ce qui se passe dans la suite du récit, pourquoi pensez-vous que l'uteur a décidé d'introduire le puma par de telles descriptions?
Finnish translation	*Back translation from Finnish*
Tässä on joitakin kertomuksen alkuvaiheessa esiintyneitä viittauksia puumaan: "…*kiljahdus* herätti hänet. Ääni oli niin tuskainen…" (rivi 37) "Vastauksena oli toistuvaa *kiljahtelua,* mutta vähemmän läpitunkevaa, väsyneen kuuloista…" (rivit 50-51) "…hän oli kuullut niiden kärsivältä kuulostavia *kiljahduksia* matkan päästä." (rivit 58-59) Ottaen huomioon, mitä kertomuksessa muuten tapahtuu, miksi arvelet kirjoittajan halunneen esitellä puuman näiden kuvausten kautta?	~Here are some of the early references to the panther in the story "the *shout* awoke her. The sound was so anguished…" (line 37) "As an answer there was a repeated *shout,* but less shrill, tired sounding…" (lines 50-51) "She had…heard their suffering-sounding *shouts* in the distance." (lines 58-59) Considering what happens in the rest of the story, why do you think the writer chooses to introduce the panther with these descriptions?

~ shows that what follows is more like an approximation (not exactly what was said, for example).

all the secondary information in their short-term memory before they got to the main point (e.g., [45, 46]).

A roughly similar albeit less significant difference also occurs in alternatives C and D. In both alternatives, the two source versions use a nonfinite relative clause (e.g., *described in the first part*), which, as is typical in the languages, follows its head (e.g., *problem*). However, in the Finnish version, in line with Finnish grammar (e.g., [43]), the nonfinite clauses precede the head. Thus, in the alternatives, too, left-branching structures are used, which, in turn, may have blurred the meanings of the two alternatives and further added to the information and memory load placed on Finnish readers, thereby making it harder for them to find the correct alternative.

The above differences were all due to language-specific differences in grammar. Therefore, not much could have been done to eliminate the bias altogether. This was the case with the question stem in particular. However, in the alternatives, another formulation would have been possible. To translate the postmodifying nonfinite relative clause, it would have been possible to use a finite relative clause, in which case the clause would have followed its head (e.g., *ongelmaa, jota kuvataan ensimmäisessä osassa* "problem that is described in the first part"), as in the English version. However, the premodifying participial structure actually used in the alternatives is a very Finnish formulation [44] and does sound a natural and fluent rendering here. Therefore, while the postmodifying finite structure would have made it possible to preserve the word order, it may perhaps have lost somewhat in idiomaticity.

5.2. Item 2. Item 2 was an open-constructed response item which occurred with a short story telling about a woman and a panther and the delicate relationship between the two during a flood in the Mississippi River. In the item, students were first presented with three direct quotations from the stimulus text, each describing the panther and its cries, and then asked why they thought the panther was introduced with these words (Table 4).

To get full credit for this full-credit/partial-credit item, students were to recognize that the descriptions were intended to evoke pity or compassion. Recognizing this, however, required that students be able to interpret nuances. This seems to have been difficult in all countries. Yet, in Finland, it was harder than expected. In the FIER [32], it was mentioned that Finnish students often got this item wrong, because they interpreted the noises made by the panther as arousing fear. However, no explanation was offered for the misinterpretation.

A semantic analysis of the Finnish and source versions shows that the misinterpretation may have been due to differences between the languages in certain key words, because of which also the nuances attached to the panther seem to have been partly different. Firstly, in the English item, each of the three references to the panther contains the word *cry*. In English, the word may refer not only to sounds of animals and shouting but also, for instance, to weeping. Moreover, especially in the sense of weeping, the word usually evokes strong connotations of grief, lamentation, pain, vulnerability, and pity. However, languages differ in their meaning structures, and there is seldom full synonymy between languages. Therefore, in the French source version,

two different words are used for the English *cry*: *hurlement* and *cri*. In the French version, the connotations thus seem to be relatively close to those in the English version. However, in the Finnish version, with *kiljahdus* "shout" and especially with its more frequentative near synonym *kiljahtelu* "shouting," most of the tender connotations of the English *cry* appear to be lacking. Therefore, in the Finnish version, the clues provided to interpret the panther seem to arouse the least pity.

Secondly, there were also lexical differences in the *unit* which may have colored readers' interpretations of the panther. One such difference had to do with the nouns used to refer to the panther (other than *panther*). The English version often uses the word *cat*, which, even though it most often denotes "domestic cat," can also refer to wild big cats. However, it seems that in both French and especially in Finnish, the closest equivalents of *cat* (*chat* and *kissa*, resp.) have somewhat more restricted meanings, in that in everyday usage they do not as readily refer to wild cats but only to domestic cats. Therefore, in both versions, slightly different nouns are employed: in the French version, the renderings are *félin*, *animal*, and *puma* and in the Finnish translation, *puuma* "panther," *kissaeläin* "feline," and *kissapeto* "feline beast." However, the connotations of these words differ. Thus, whereas the English *cat* paints a picture of a cute, sweet, cuddly, disarming, and harmless pet, with the French *félin*, *animal*, and *puma*, the picture is clearly more neutral and less petlike. However, with the Finnish "panther," and "feline beast," in particular, the picture seems even dangerous and frightening. This may also have made some Finnish students associate the panther, not with the feeling of pity but fear. This assumption seems to be supported by the fact that also the other two items in this unit where the term *cat* played a significant role proved unusually difficult in Finland [29].

Even though both the above differences in the key words basically arose from language-specific differences in meaning, slight modifications could perhaps have been made to the Finnish version to render it more equivalent to the source versions. In place of "shout," one could perhaps have used, for example, *ulvonta* "howling, wailing," with which the connotations of weeping and pain would have been somewhat closer to those in the source versions. In the case of *cat*, the more neutral equivalents (e.g., "feline") could have been favored over the more threatening "panther" and, especially, "feline beast." However, while these revisions would probably have made the Finnish version more equivalent to the French version in particular, the English version would still have stood out as the most different and the least equivalent.

6. Results and Discussion

This study examined whether it would be possible to improve the ability of judgmental reviews conducted in international education studies to accurately and reliably identify language-related sources of bias. A new review was made of two PISA 2000 items which in Finland showed DIF but for which no language-related explanations were found in 2000 or for which the explanations were unclear. The items were compared linguistically to the corresponding source-text items, and students' written responses to the items were consulted to gain insight into their thought processes. In the review, both Finnish items were found to differ linguistically from the source items in a way that might have affected students' performance on the items. In both cases, the differences were basically due to differences between languages, either in grammar and syntax or in meanings and connotations. However, in at least one of the items, the differences also seemed to be partly attributable to the translation not having been fully appropriate. The study thus suggests that it is possible to make judgmental reviews better able to identify language-related sources of bias, especially when they derive from differences between languages but also when they are the result of inappropriate translation. However, to reach this goal, a more standardized judgmental review procedure needs to be developed (see also [9, 10, 47]). In the following, suggestions are given on how this may be done.

6.1. Developing Judgmental Reviews. First, a clear definition should be provided as to what is meant by "translation problems." At the moment, it seems that they are mainly understood as involving only more or less outright translation errors, whereas it has been much more difficult to decide what to do with more subtle inaccuracies resulting from differences between languages, for example. During the first PISA round, reviewers were still encouraged to look also for language-specific differences [30], but in more recent studies, no mention has been made of them (e.g., [31]). However, there is ample research suggesting that differences between languages often lead to bias (e.g., [7, 9, 15]). Therefore, it is important to include also these differences in the definition of translation problems (see also [13]). Also, more information and examples are needed on the different types of translation problems and the most typical sources of language-related bias (see also [7]).

Second, instructions should be provided on how to conduct the review. Today, such instructions are largely missing. The review could be conducted as in this study (see Table 2 and Section 5). It should involve a thorough and systematic linguistic comparison of the target and source versions against the reading strategies assessed. The comparisons should be made between the DIF items *and* (the relevant parts of) the stimulus texts with which they occur (as shown by *cat* in Item 2), and they should concentrate on the typical sources of language-related bias (see Table 1; cf. [13, page 82]) and factors affecting item and text difficulty. Reviewers should also consider their judgments against testees' written responses, as revealed in national review reports (e.g., [32]), for example. This would provide them with a rough, tentative glimpse into the cognitive processes of the testees, thereby making it possible for them to make stronger and more reliable hypotheses about the sources of DIF (cf. [11]).

Third, the instructions on who, and with what qualifications, should conduct the review need to be clarified (see also [14]). Expertise in the domain assessed and familiarity with test development, as required today, are not enough.

Knowledge of reading processes, response processes, and factors affecting item difficulty are, of course, needed. In addition to this, however, to be able to compare the difficulty of the items across languages and to detect all the at times subtle translation problems and linguistic differences between the versions, the reviewer also needs linguistic and translation expertise and a good command not only of the target language but also of the source languages. Finding one reviewer who would fulfill all the requirements may be difficult, and therefore it may often be necessary to have at least two reviewers. Having several reviewers would also add to the reliability of the judgments: given the great number of factors (e.g., linguistic levels) that have to be taken into account when making the judgments, using several reviewers would help to spot more potential problems. Also, if several reviewers were used, one or some of them could avail themselves of the data on how the items functioned in the country and how students answered them while other(s) could work blinded.

Fourth, instructions should be provided on how to appraise and rate the significance of language-related differences on performance (e.g., minor, moderate, major), because these vary greatly in gravity (e.g., the difference in alternatives C and D in Item 1 seemed less significant than the other differences in this study; see also [13]). No such instructions exist today. A verbal or numerical rating scale could be developed for this purpose (cf. [4, 10, 36, 48]). Such ratings can be a great help to the Project Centre, which cannot have expertise in all languages and is therefore largely dependent on the national reviews, when making the final judgmental decisions.

Fifth, the results of the national reviews need to be reported in an unequivocal manner to the Project Centre so that they can make justified decisions on the items. To this end, it is important to refine the sheet provided for this purpose so that it contains, for each flagged item, a clear explanation as to why the reviewers think it functions poorly and an appraisal as to how significant they think the difference is (cf. [40]).

Sixth, a brief instruction manual needs to be prepared on judgmental reviews, which should contain, among others, the definition and examples of translation problems, a description of the method for conducting the review, information and examples regarding typical sources of language-related bias, instructions on how to rate the significance of the differences, and the requirements set for the reviewers. Each reviewer should get a copy of the manual. At the moment, no such manual exists (except for the small section devoted to judgmental reviews in NPM manuals, which only contains very little information, is intended for NPMs, who may not be directly involved in the review, and is supplied long before the reviews are conducted, because of which the information easily gets lost and is forgotten). In addition to being provided with the manual, the reviewers also need training on how to conduct the review.

6.2. Developing Other Quality Control Procedures. On the basis of the sources of DIF found in this study, development also seems to be needed in other quality control procedures followed in international education studies.

6.2.1. Item Generation. As revealed by this study and several other studies (e.g., [7, 9, 15]), differences in grammatical form and structure and differences in meanings and connotations are a common potential source of bias in international education studies. Moreover, contrary to the view that seems to have prevailed in PISA 2000, that language-specific differences are repairable [30], often not much can be done to correct these differences and to eliminate the bias [13, 40]. The fact is that languages and their ways of conveying meaning are different, and therefore all attempts to by force make different-language test versions "similar" easily result in the translations being cumbersome and difficult to understand (e.g., [49])—and, as a result, nonequivalent to the idiomatic and natural source versions. The difficulty of completely eliminating language-related bias was also seen in this study, where, even though slight revisions would sometimes have been possible and even in order, the revisions would not have made the items fully equivalent. Therefore, often the only way to remove bias caused by a language-specific difference is to delete the item.

However, item deletions are never the best solution, because they reduce content and cognitive domain coverage, thereby lowering the reliability and validity of test scores, and are also a waste of time and resources. Therefore, it is imperative to try to take these sources of bias into consideration *a priori*, while still generating the items. A good deal of information already exists on typical sources of language-related bias, which can be used when generating items. However, it is good to continue collecting more such data during each PISA round and also from other sources (see also [7, 15]). For example, from this study one can see that because of language-specific differences in form, items which are based on and ask about the exact wording of an expression (like Item 1) are always risky (see also [11]). Similarly, since languages differ greatly in their meaning structures, also items drawing on connotations and nuances (Item 2) seldom function equivalently across languages. Another efficient help in item generation is to involve translators competent in various languages in the process (see also [6, 50]), even more than is done today. Actually translating the items into several, possibly remote languages (not only into French) helps to see what does not work and why and to modify the items accordingly.

6.2.2. Translation. However, while completely eliminating bias caused by language-specific differences is usually not possible, slight revisions can often be made to reduce the bias (see also [7, 13]). This was the case also in this study in Item 2 in particular, where the bias would probably have decreased, if slightly different translations had been used for the two key words, *cry* and *cat*. This suggests that the translations were not as equivalent as they could have been and that more equivalent translations could have been made, if translators, while translating, had paid more attention to the key words (see also [10, 11]) and their connotations. This emphasizes the need for good translators and specific training for them

(see also [6]). Also, special translation notes can be used to remind translators of the importance of the key words (see also [7]). In more recent PISA studies, more attention has been paid to this use of translation notes.

7. Limitations of the Study and Suggestions for Further Research

When interpreting the findings of the study, it should be remembered that the study had its limitations. First, even though the consideration of testees' written responses may have helped to make more informed and reliable hypotheses about testees' thought processes and sources of DIF, it could not provide direct and explicit data on the processes, only indirect and tentative hypotheses. The findings are thus still hypotheses and need to be validated by empirical investigations, such as TAPs and cognitive interviews. However, this is also true of all *a posteriori* judgmental reviews [11]. Yet, conducting such investigations (on all items, countries and languages) is a huge and expensive undertaking. Therefore, future research could examine how reliable testees' written responses really are (or can be made to be) in identifying sources of bias and whether these responses could perhaps, under certain circumstances (i.e., in certain types of open-ended items; cf. Item 2), be sufficient to reliably decide whether items are biased. This, of course, further requires that countries carefully analyze and document data on testees' written responses (e.g., in national review reports).

The second limitation of the study is that only two items used in the PISA 2000 reading test and translated into Finnish were analyzed in this study. Comparable reviews of larger numbers of DIF items, of versions translated into other (e.g., Asian) languages, and of other international tests are, of course, needed to corroborate the findings and to justify the recommendations made in this study. However, previous research on factors affecting text difficulty and the typical sources of bias suggest that the problems discussed in this study—grammatical and semantic differences between languages (e.g., word order, connotations, transparency, processing load)—are universal and that the findings have a wider application.

Studies with greater numbers of items are needed also for other reasons. The fact that in this study, language-related explanations were found for both the DIF items for which such explanations were not clearly identified in 2000, and that *cat* even seems to have been a problem in two more items, suggest that the true amount of language-related bias in Finland may have been higher than judged in 2000. Moreover, considering the deficient review reports of several participating countries and the great number of countries which, like Finland, had no or few national deletions, this may have been true of several other countries, too. If this is the case, it may threaten the validity of the PISA 2000 reading test. Furthermore, since items developed in PISA 2000 have been used as anchoring items in later PISA rounds, doubts are cast on the validity of these studies, too. Studies with greater numbers of items are thus needed to decide on the validity of the tests. Such studies will also help to collect more information on the typical sources of language-related bias and to see whether these vary across text types, item formats, and aspects, for example. All this information will be a great help in test development.

At the same time, however, it is important to remember that there are limits to what quality control procedures, including judgmental reviews, can do. As shown by also this study (Item 1 in particular), languages are different and convey meaning in different ways, and sometimes this inevitably leads to differences, for example, in processing and cognitive load and, in the end, in difficulty. Therefore, nothing will ever ensure fully equivalent translations. This has led some researchers (e.g., [51]) to suggest that countries use indigenous, untranslated texts and items in these studies and that the comparability of these materials be ensured by analyzing them against a given set of criteria of text and item difficulty. This would help to ensure greater idiomaticy and cultural fairness. Future studies could examine whether it would be possible to combine the strengths of these two approaches (translated versus untranslated texts).

Acknowledgment

The preparation of this paper has been supported by the Academy of Finland (Grant no. 126855).

References

[1] J. Sweller, "Cognitive load during problem solving: effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[2] R. Rueda, "Cultural perspectives in reading: theory and research," in *Handbook of Reading Research*, M. Kamil, P. D. Pearson, E.B. Moje, and P. Afflerbach, Eds., vol. 4, pp. 84–104, Routledge, New York, NY, USA, 2011.

[3] R. Tourangeau, L. Rips, and K. Rasinski, *The Psychology of Survey Response*, Cambridge University Press, Cambridge, Mass, USA, 2000.

[4] K. Ercikan, M. J. Gierl, T. McCreith, G. Puhan, and K. Koh, "Comparability of bilingual versions of assessments: sources of incomparability of English and French versions of Canada's national achievement tests," *Applied Measurement in Education*, vol. 17, no. 3, pp. 301–321, 2004.

[5] F. Guérin-Pace and A. Blum, "The comparative illusion: the international adult literacy survey," *Population*, vol. 12, pp. 215–246, 2000.

[6] R. Hambleton, "Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, R. Hambleton, P. Merenda, and C. Spielberger, Eds., pp. 3–38, Erlbaum, Mahwah, NJ, USA, 2005.

[7] A. Allalouf, "Revising translated differential item functioning items as a tool for improving cross-lingual assessment," *Applied Measurement in Education*, vol. 16, no. 1, pp. 55–73, 2003.

[8] I. Arffman, "The problem of equivalence in translating texts in international reading literacy studies. A text analytic study of three English and Finnish texts used in the PISA 2000 reading test," Research Reports 21, University of Jyväskylä, Institute for Educational Research, Jyväskylä, Finland, 2007.

[9] P. Elosua and A. López-Jaúregui, "Potential sources of differential item functioning in the adaptation of tests," *International Journal of Testing*, vol. 7, no. 1, pp. 39–52, 2007.

[10] K. Ercikan, "Disentangling sources of differential item functioning in multilanguage assessments," *International Journal of Testing*, vol. 2, no. 3-4, pp. 199–215, 2002.

[11] K. Ercikan, R. Arim, D. Law, J. Domene, F. Gagnon, and S. Lacroix, "Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews," *Educational Measurement*, vol. 29, no. 2, pp. 24–35, 2010.

[12] A. Grisay, "*PISA 2000: differences in item difficulty between English, French and German countries*," Unpublished manuscript, 2004.

[13] G. Solano-Flores, E. Backhoff, and L. Contreras-Niño, "Theory of test translation error," *International Journal of Testing*, vol. 9, no. 2, pp. 78–91, 2009.

[14] G. Engelhard, L. Hansche, and K. Rutledge, "Accuracy of bias review judges in identifying differential item functioning on teacher certification tests," *Applied Measurement in Educatio*, vol. 3, no. 4, pp. 347–360, 1990.

[15] M. J. Gierl and S. N. Khaliq, "Identifying sources of differential item and bundle functioning on translated achievement tests: a confirmatory analysis," *Journal of Educational Measurement*, vol. 38, no. 2, pp. 164–187, 2001.

[16] L. Roussos and W. Stout, "A multidimensionality-based DIF analysis paradigm," *Applied Psychological Measurement*, vol. 20, no. 4, pp. 355–371, 1996.

[17] A. Grisay and C. Monseur, "Measuring the equivalence of item difficulty in the various versions of an international test," *Studies in Educational Evaluation*, vol. 33, no. 1, pp. 69–86, 2007.

[18] A. Grisay, E. Gonzalez, and C. Monseur, "Equivalence of item difficulties across national versions in PIRLS and PISA reading assessments," *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, vol. 2, pp. 63–84, 2009.

[19] A. Grisay, H. de Jong, E. Gebhardt, A. Berezner, and B. Halleux-Monseur, "Translation equivalence across PISA countries," *Journal of Applied Measurement*, vol. 8, no. 3, pp. 249–266, 2007.

[20] R. Adams and M. Wu, Eds., PISA 2000 Technical Report, OECD, Paris, France, 2002.

[21] R. Adams, "Scaling PISA cognitive data," PISA 2000 Technical Report, OECD, Paris, France, 2002, Edited by R. Adams, M. Wu.

[22] J. McQueen and J. Mendelovits, "PISA reading: cultural equivalence in a cross-cultural study," *Language Testing*, vol. 20, pp. 208–224, 2003.

[23] R. Adams and C. Carstensen, "Scaling outcomes," PISA 2000 Technical Report, OECD, Paris, France, 2002, Edited by R. Adams, M. Wu.

[24] OECD, *Translation, Adaptation and Verification of Test Material in OECD International Surveys*, OECD, Paris, France, 2010.

[25] J. Olson, M. Martin, and I. Mullis, Eds., *TIMSS 2007 Technical Report*, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, Mass, USA, 2008.

[26] OECD, *Reading for Change. Performance and Engagement Across Countries. Results from PISA 2000*, OECD, Paris, France, 2002.

[27] OECD, PISA 2006 Technical Report, OECD, Paris, France, 2009.

[28] OECD, *National Project Manager's Manual*, OECD, Paris, France, 1999.

[29] OECD, Report on the dodgy items detected in Finland in the PISA 2000 reading literacy assessment, Unpublished raw data, 2000.

[30] OECD, *National Project Manager's Manual*, OECD, Paris, France, 2000.

[31] OECD, *NPM Manual PISA 2009. National Project Managers' Manual for the PISA 2009 Survey*, OECD, Paris, France, 2009.

[32] Finnish Institute for Educational Research, Finland's report on the judgmental review of the dodgy items in the PISA 2000 reading literacy assessment, Unpublished raw data, 2000.

[33] S. M. Corey, *Action Research to Improve School Practices*, Teachers College Press, New York, NY, USA, 1953.

[34] J. McKernan, *Curriculum Action Research*, Kogan Page, London, UK, 2nd edition, 2000.

[35] M.Q. Patton, *Qualitative Research & Evaluation Methods*, Sage, Thousand Oaks, Calif, USA, 3rd edition, 2002.

[36] I. S. Kirsch, *The International Adult Literacy Survey (IALS): Understanding what was Measured*, Educational Testing Service, Princeton, NJ, USA, 2001.

[37] P. B. Mosenthal and I. S. Kirsch, "Toward an explanatory model of document literacy," *Discourse Processes*, vol. 14, pp. 147–180, 1991.

[38] C. Alderson, *Assessing Reading*, Cambridge University Press, Cambridge, UK, 2000.

[39] J. Chall and E. Dale, *Readability Revisited: The New Chall-Dale Readability Formula*, Brookline Books, Cambridge, Mass, USA, 1995.

[40] A. Allalouf, R. Hambleton, and S. Sireci, "Identifying the causes of dif in translated verbal items," *Journal of Educational Measurement*, vol. 36, no. 3, pp. 185–198, 1999.

[41] M. A. K. Halliday, *An Introduction to Functional Grammar*, Arnold, London, UK, 2nd edition, 1994.

[42] A. Baugh and T. Cable, *A History of the English Language*, Routledge, London, UK, 2002.

[43] F. Karlsson, *Finnish: An Essential Grammar*, Routledge, London, UK, 1999.

[44] R. Ingo, *Suomen kieli vieraan silmin*, University of Vaasa, Vaasa, Finland, 2000, [The Finnish language through the eyes of a foreigner] (Research Group for LSP and Theory of Translation, No. 26).

[45] R. C. Anderson and A. Davison, "Conceptual and empirical bases of readability formulas," in *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, A. Davison and G. Green, Eds., pp. 23–53, Erlbaum, Hillsdale, NJ, USA, 1988.

[46] I. Schlesinger, *Sentence Structure and the Reading Process*, Mouton, The Hague The Netherlands, 1968.

[47] H. Yildirim and G. Berberoğlu, "Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items," *International Journal of Testing*, vol. 9, pp. 108–121, 2009.

[48] C. Jeanrie and R. Bertrand, "Translating tests with the international test commission's guidelines: keeping validity in mind," *European Journal of Psychological Assessment*, vol. 15, no. 3, pp. 277–283, 1999.

[49] E. Nida and C. Taber, *The Theory and Practice of Translation*, Brill, Leiden, The Netherlands, 1969.

[50] O. Behling and K. Law, *Translating Questionnaires and Other Research Instruments: Problems and Solutions*, Sage, Thousand Oaks, Calif, USA, 2000, (Sage University Papers Series on Quantitative Applications in the Social Sciences, no. 07–133).

[51] G. Bonnet, F. Daems, C. de Clopper et al., *Culturally Balanced Assessment of Reading (c-bar). A European Project*, European Network of Policy Makers for the Evaluation of Education Systems, Paris, France, 2003.