PERGAMON

International Journal of
Educational Research 29 (1998) 543–553

International Journal of
Educational
Research

Chapter 4

Translation effects in international assessments

Kadriye Ercikan

University of British Columbia, Canada

Abstract

Ideally, a single common form of a test would be used for international assessments. However, since the test is administered in different countries, it is necessary to translate the test into the languages of these countries. This chapter explores the application of a statistical method to examine the effect of translations on the equivalence of test items and the comparability of test scores. This method is used to identify poorly translated items in an international assessment which was administered in two languages and to examine how the comparability of scores is affected by problems in translations.

Keywords: International assessment; Content validity; TIMSS; Translation effects; Sampling effects; Score level; Mathematics and science assessment; International comparisons

International assessments are being increasingly valued by ministries of education in different countries. These assessments allow comparisons of educational input, process, and achievement in participating countries and can provide a broad perspective for evaluating and improving education. The complex nature of international assessments makes them very sensitive to the methodologies used and the validity of such comparisons depends on these methodologies. Equivalence or comparability is a central issue in discussions on the methodology of international assessments. Equivalence or comparability is concerned with whether some common construct can be inferred from measurements of different groups of students in different countries. The focus of this chapter is the effect of translations on equivalence of tests and comparability of scores in international assessments.

Ideally, a single common form of a test would be used for international assessments. However, since the test is administered in different countries, it is necessary to translate the test into the languages of these countries. If done improperly, translations can influence psychometric properties of tests. The quality of a translation affects its accuracy in terms of meaning, connotations, style, and degree of difficulty of key vocabulary and passages of items. A poor translation can result in misleading or confusing language which interferes with the students' ability to comprehend test items and respond to them (Brislin, 1988; Cabello, 1983; Brislin et al., 1973; Hulin et al., 1983).

0883–0355/98/$19.00
PII: S0883-0355(98)00047-0

A good translation reflects the meaning, intent, tone, and general style of the original version (Rodrigues-Bou, 1956). It must reflect not only the meaning of the original item, but should also try to maintain the same relevance, intrinsic interest, and familiarity of the item content, otherwise what the item measures may be altered.

There are several basic differences in languages which cause problems in translations. Among these are the variations in the frequency of word use and in word difficulty. Words which may be commonplace and "easy" in one language may not be equally so in another. Another translation problem occurs when grammatical forms either do not have equivalents, or else have many of them in one or the other language. Syntactical style is one of the most difficult features to carry over from one language to another.

Translations as potential sources of bias can affect the meaning and functions of single words, sentences, and passages, the content of the items, and the skills measured by the items. The degree and manner in which item features are changed during translation will determine whether the equivalence of items is maintained. Changes in any of these item features may alter its difficulty or even what is being measured.

To assess both translation quality and equivalence of source and target versions, researchers have dealt with: (1) comparisons of meaning, (2) the ability to gain the same information from reading both source and target versions, (3) responses to different language versions, and (4) performance measures. Brislin et al. (1973) present the following guidelines to help others write translatable English:

(1) Use short, simple sentences of fewer than 16 words.
(2) Employ the active rather than the passive voice.
(3) Repeat nouns instead of using pronouns.
(4) Avoid metaphors or colloquialisms. Such phrases are the least likely to have equivalents in the target language.
(5) Avoid the subjunctive mode, for example, verb forms with "could" or "would".
(6) Avoid adverbs and prepositions telling "where" or "when" (e.g., "frequent", "beyond", "upper").
(7) Avoid possessive forms where possible.
(8) Use specific rather than general terms (e.g., the specific animal, such as cows, chickens, pigs, rather than the general term "livestock").
(9) Avoid words which indicate vagueness regarding some event or thing (e.g. "probably" and "frequently").
(10) Avoid sentences with two different verbs if the verbs suggest different actions.

This chapter explores the application of a statistical method to examine the effect of translations on the equivalence of test items and the comparability of test scores. This method is used to identify poorly translated items in an international assessment which was administered in two languages and to examine how the comparability of scores is affected by problems in translations.

1. Detection of translation effects

When items are poorly translated, their properties may change for the groups taking the test in different languages. These changes in properties of items can affect

what is being assessed by the test as well as altering the difficulty of the item for different groups. When an item displays varying properties in different group settings, after controlling for differences in the abilities of the groups, these variations in properties of items are called *differential item functioning* (DIF) (Angoff, 1993). In international assessments, translations are one of the factors that can affect properties of items in different countries and cause DIF. Therefore, statistical methods used to detect DIF can be used to identify poorly translated items.

One way of interpreting DIF, which is sometimes referred to as item bias, is described by Ackerman (1992). "If two different groups of examinees have different underlying multidimensional ability distributions and the test items are capable of discriminating among levels of abilities on these multiple dimensions, then any unidimensional scoring scheme has the potential to produce bias" (p. 67). In the case of international assessments, multidimensional abilities can differ from one country to another due to cultural, language, and curriculum differences. Cultural differences can influence intrinsic interest and familiarity of the content of items. Curriculum-related differences can result in varying degrees of student exposure to the domain of items. Certain topics or subjects will undoubtedly have differential coverage in different countries and differential coverage can lead to differential response patterns and difficulty levels, independent of any problems due to translation. When tests are assumed to be unidimensional and these differences are not taken into account, the results based on a single dimension can give rise to DIF. Given all possible causes of DIF, items statistically identified as showing DIF are not necessarily poorly translated items; however, they are good candidates for investigating potential translation problems.

2. Example

The utility of using DIF procedures to examine the equivalence of test items and comparability of scores from tests in different languages was explored using the 1984 International Association for the Evaluation of Educational Achievement (IEA) science tests in English and French. The IEA test used in that study was constructed by an international committee of educational experts covering material that they believed should have been mastered by children of a certain age or grade group. In the development of the science test, a group of scientists and science educators determined those aspects of science that students in a particular grade should know if they are to become good scientists. If, for example, a particular science concept was considered to be important, the committee decided to include the concept even though it was taught in only 5 out of 20 countries. Once the items were written in the source language, in this case English, they were translated to the target language(s), in which the non-English-speaking students were being tested. This involved (a) the translation of items from the source language to the target language, (b) the translation of these back into the source language, and (c) the comparison of the two versions of items in the source language and making the right adjustments to both versions.[1] For the purposes

[1] Personal communication with T. N. Postlethwaite, April 1989.

of this study, 70 common items in English and French versions of the test were used.

Assessment data from the 1984 IEA Science Study (Population 2, 14 yr old) on two groups, namely the Canadian English- and French-speaking students were used. These two groups are from the same country and are products of the same educational system. Therefore their educational characteristics are expected to be more similar than countries typically found in international assessments. The English-speaking group consisted of 5543 students and the French-speaking group consisted of 2348 students.

The analyses involved first using a DIF detection procedure to flag items that may potentially have translation problems. Items that are flagged as manifesting DIF are then studied closely in the languages of both comparison groups to identify translation problems. Finally, the impact of non-equivalence of items detected by the DIF procedure on comparability of scores is examined.

2.1. DIF analyses

Several different statistical procedures have been used to detect differential response patterns to items by groups of examinees who have the same ability levels. Holland (1985) proposed the use of the Mantel-Haenzsel procedure as a relatively inexpensive yet statistically powerful technique for identifying differentially functioning test items. This method, was employed in this study. The statistical analyses involved studying each item separately. A number correct score was produced based on the 70 items. The responses to the target item were compared, controlling for achievement level as measured by the total score on the test. This process was repeated for each item and the items showing DIF, favoring one group or the other, were identified. The items which were omitted or not reached by the examinees were treated as incorrect.

The Mantel-Haenzsel method involves the creation of J two-by-two tables, where J is the number of score categories. For the jth score level, the data can be displayed as in Table 1.

Here, F denotes the focal group (Black, Hispanic, or other ethnic groups, or females in general; in this study, the French-speaking group) and R denotes the reference group (Whites or males in general; the English-speaking group here). The numbers of examinees in the R and F groups are denoted by n_{Rj} and n_{Fj}, respectively; m_{1j} represents the number of examinees who answered the item correctly and m_{0j} is

Table 1
Data for the jth matched set of reference and focal group members

Group	Score on studied item		Total
	1	0	
Reference	A_j	B_j	n_{Rj}
Focal	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

the number who answered incorrectly. A_j and C_j denote the numbers of examinees in the R and F groups, respectively, who answered correctly; B_j and D_j are the numbers of examinees in the R and F groups who answered incorrectly. T_j is the total number of examinees.

As described in Holland and Thayer (1988), it is assumed that, within each stratum, data for the R and F groups have been acquired by obtaining (simple) random samples of fixed sizes (n_{Rj} and n_{Fj}) from pools of reference and focal group members. A_j and C_j are then independent binomial random variables with parameters (n_{Rj}, p_{Rj}) and (n_{Fj}, p_{Fj}), respectively. In the present context, P_{Rj} represents the probability of answering the item correctly for members of the reference group in the jth stratum; P_{Fj} is the corresponding probability for the focal group. Then the hypotheses to be tested are as follows:

$$H_0: \quad (P_{Rj}/Q_{Rj})/(P_{Fj}/Q_{Fj}) = 1, \quad j = 1, \ldots, J, \text{ where } Q_{Rj} = 1 - P_{Rj},$$

versus

$$H_1: \quad (P_{Rj}/Q_{Rj})/(P_{Fj}/Q_{Fj}) = \alpha, \quad \alpha \neq 1.$$

The parameter α represents the common odds ratio for the J 2×2 tables and indicates the degree of DIF for each item. In Mantel-Haenzsel procedure the statistic typically used as an index of DIF is

$$\text{MH D-DIF} = -2.35 \ln(\alpha) \quad \text{(see Holland and Thayer, 1988).}$$

This method allows us to categorize items according to their degree of differential functioning for different groups. Items are categorized as "DIF Free" if MH D-DIF is not significantly different from 0 or has an absolute value less than 1. Items are categorized as "Low DIF" if MH D-DIF is significantly different from 0 and has either (a) an absolute value at least 1 but less than 1.5 or (b) an absolute value at least 1 but not significantly greater than 1. Items are categorized as "High DIF" if MH D-DIF is at least 1.5 and is significantly greater than 1. These classifications correspond to those used by other researchers who used Mantel-Haenzsel procedure to examine DIF items (Zwick and Ercikan, 1989).

2.2. Equivalence of items in French and English

Table 2 displays some descriptive statistics of the IEA Science test administered in English and French. These statistics indicate small differences in the performance of English and French speaking students as well as the reliability of the test for the two groups. The average percent correct score for the English-speaking group was 42.9 and the internal consistency coefficient alpha was 0.78. The average percent correct score for the French-speaking group was 41.0 and the internal consistency coefficient alpha was 0.74. Both the average percent correct score and coefficient alpha were slightly lower for the French-speaking group. However, none of these differences are sufficient to indicate non-equivalence between the versions of the test in English and French. Both the percent correct statistic and the reliability statistic are conditional

on the science ability distribution which is not expected to be the same in these two groups.

DIF analyses, on the other hand, condition on ability level and, therefore, provide information about whether the items are assessing similar constructs for the two groups. Table 3 presents the numbers of items identified as manifesting DIF in favor of the reference group or the focal group. The analysis revealed that there were some items on which focal group students performed worse than reference group students, and vice versa, conditional on the number correct score. The statistical analysis identified 18 items (26% of the total test) as showing DIF. Eight were in favor of the French-speaking group and ten were in favor of the English-speaking group. Two of those items in favor of the English-speaking group were classified as *Low DIF* and eight were classified as *High DIF*. Four of the items in favor of the French-speaking group were classified as *Low DIF* and four were classified as *High DIF* items. These results indicate that, overall, there was stronger differential item functioning in favor of the English-speaking reference group.

Items classified as *Low DIF* or *High DIF*, and as showing DIF in favor of one group or another, were examined and compared in the two languages. For these items, the correct response rates differed and some of the distractors that were functional for one group were not functional for the other group. To examine probable causes of these differences, these items were further examined with the help of several translators. The following is a summary of the observations made.

1. In one item, a particular word was determined to be in more common usage in French than in English. In English "the animal preyed on other animals" was translated into French as "the animal fed on other animals". The fact that "prey" is a less common word then "fed" could have made the item more ambiguous and difficult for the English-speaking group.

Table 2
IEA science assessment descriptive statistics

Group	Sample size	Coefficient alpha (70 items)	Average percent correct
Reference	5543	0.78	42.9
Focal	2348	0.74	41.0

Table 3
IEA results of DIF analysis. Numbers of *Low DIF* and *High DIF* items

In favor of reference group		In favor of focal group	
Low DIF	High DIF	Low DIF	High DIF
2	8	4	4

2. An item was longer and was phrased in more complicated sentences in one language than the other, thus possibly making the question harder to understand.
3. An item that appeared in a single sentence in English consisted of two sentences when translated to French. In the French test, the actual question was separated from the data provided thus possibly making it easier to understand.
4. The key word in an item had a broader meaning in French than English. The word "abdomen", when directly translated to French, has a broader meaning. This difference in meaning could have led to the misinterpretation of the item.
5. In English, the distractor "There is no air on the Moon to offer resistance" was translated to French as "There is no air on the Moon which offers resistance". The two sentences have slightly different meanings. The resulting ambiguity could have caused differential responses to this item.
6. The word "moist" in a distractor was translated as "humid" in French. "Moist" is not as common a word as "humid" is for describing weather. The distractor containing "moist" was avoided more frequently by the English speaking group.
7. In English, a word had more than one meaning. An item containing physics terms like "energy" and "power", also had the word "work" in one of the distractors. "Work" in English has a meaning outside the context of physics as well. When directly translated to French, this word did not have the physics context meaning. For the French group, the distractor involving "work" was not functional and it might have caused the item to be easier.
8. The item that showed the highest DIF (displayed in Table 4 as item #1) was about the function of human perspiration. The correct answer was its cooling effect. The word "cool" was translated as "refresh" in French. A greater proportion of the English-speaking group (45%) had answered this item correctly, whereas only 10% of the French group answered it correctly.

In summary, for eight out of the 18 (44%) DIF items had interpretations related to translation problems. Additionally, these eight items violated the rules for writing

Table 4
Sample items

Item number	Item content
1	The formula for the acetic acid (present in vinegar) is CH_3COOH. What is the total number of atoms in one molecule of acetic acid? A. 1 B. 2 C. 3 D. 6 E. 8
2	What is the main way that sweating helps your body? A. It cools your body. B. It keeps your skin moist. C. It keeps you from catching cold. D. It gets rid of excess water in your body.

translatable English offered by Brislin et al. (1973). Five of the items had sentences with more than 16 words, one item used passive voice, one item used a pronoun, and one item had two verbs which suggested different actions.

For ten of the 18 items, DIF could not be related to translation problems. An example of each of the two types of items identified as showing DIF, one with and one without interpretations related to translation problems, is displayed in Table 4. Item #1, with a possible translation problem, had the highest DIF. Item #2, with the second highest DIF in the test, had no explanation related to translation problems. It is not surprising that not all the DIF items could be explained in terms of translation problems. As discussed earlier, previous research indicates that other differences between the comparison groups, such as cultural and curricular, could lead to DIF as well.

2.3. Equivalence of test items in French and English in Another International Assessment

In order to compare the occurrence of DIF in IEA research to other similar international assessments, a study conducted by the Educational Testing Service (ETS) was examined. ETS conducted an international assessment study called the International Assessment of Educational Progress (IAEP) in February 1988. Mathematics and science achievement of representative samples of 13 yr-old from five countries and four Canadian provinces (ETS, 1988) were assessed. For translation of items into different languages, ETS followed a process similar to that used by the IEA study and used the Mantel–Haenzsel method for identifying DIF items. One of the DIF analyses of the IAEP compared a French-speaking Quebec population as the focal group and English-speaking American students as the reference group. The statistical analysis identified 28 out of 60 items (47%) as showing DIF, considerably higher percentage than the IEA study (26%). Eleven were in favor of the French-speaking group and 17 were in favor of the English-speaking group. Six of those in favor of the English-speaking group were classified as *Low DIF* and 5 were classified as *High DIF* items. These findings indicate that there was stronger differential functioning in favor of the English-speaking group, similar to the IEA study.

The difference in the proportion of DIF items in the two studies can be attributed to the different comparison groups and the approaches used to construct the tests. The DIF analysis conducted using the IEA data compared English- and French-speaking Canadian students, whereas the results from the IAEP study are based on analysis that compared Canadian students from Quebec and students from the U.S.A. The groups compared in the IEA study are expected to have more similar curricula and

Table 5
IAEP results of DIF analysis. Numbers of *Low DIF* and *High DIF* Items

In favor of reference group		In favor of focal group	
Low DIF	High DIF	Low DIF	High DIF
6	11	6	5

fewer cultural differences. As far as the test construction processes for the two studies are concerned, the domain of items in the IEA study was specified based on a consensus process among all the countries involved. In the IAEP study, on the other hand, the items were selected from an item bank created for the USA students. The potential inappropriateness of items for the Canadian students could have increased the impact of curriculum-related differences.

The higher percentage of DIF items in the IAEP study may be attributable in part to either or both of these factors. Another factor could be the differential quality of translation in the two studies. For conclusive comparisons, further analyses studying DIF items in IAEP, in English and French, would need to be conducted.

3. Impact of DIF on comparability of scores

In the interpretation of the results of DIF analyses, an important issue is the impact of DIF on the comparability of scores from different test versions of an international assessment. One way of examining this is to look at the impact of DIF in terms of differences in item difficulties (item *p*-values) for the two groups conditional on the ability level. The Mantel-Haenzsel (MH-DIF) (Holland and Thayer, 1988) statistic is given in terms of an item difficulty metric by Holland and Wainer (1993). Using their definitions, differences in item *p*-values due to DIF with MH-DIF = 1.5 can be computed as follows:

$$P_r = 0.528P_f/((1 - P_f) + 0.528\mathrm{P}_f),$$

where P_r is the *p*-value for the reference group and P_f is the *p*-value for the focal group, is the equation for DIF in favor of the focal group and

$$P_r = 1.893\ P_f/((1 - P_f) + 1.893P_f)$$

is the equation for DIF in favor of the reference group.

For item *p*-values ranging from 0.1 to 0.9 for the focal group, the estimated *p*-values for the reference group and the differences between the two groups are presented in Table 6. For these *p*-values, the differences range from 0.04 to 0.16.

A MH-DIF value of 1.5 is the minimum for an item to be classified as *High DIF*, which means that all *High DIF* items have MH-DIF 1.5 or greater. This indicates that differences seen on Table 6 are among the low end of the differences in *p*-values that can be expected. In the IEA study, eight items were classified as *High DIF* in favor of the English-speaking group and four were classified as *High DIF* in favor of the French-speaking group. In this case, in terms of the number-correct scores, DIF can lead to 0.32 to 1.28 number correct score points in favor of the English-speaking group and 0.16 to 0.64 number correct score points in favor of the French-speaking group. Even though these differences seem very small on a test with 70 items, they can lead to different rankings of countries in international comparisons. The number of DIF items is expected to be higher for groups with more differing educational and cultural characteristics than the English- and French-speaking Canadian groups studied in this study.

Table 6
Differences in item *p*-values due to DIF. In favor of focal group in favor of reference group

Focal Group	Reference Group	Difference	Focal Group	Reference Group	Difference
0.10	0.05	0.04	0.10	0.17	0.07
0.20	0.12	0.08	0.20	0.32	0.12
0.30	0.18	0.12	0.30	0.54	0.15
0.40	0.26	0.14	0.40	0.56	0.16
0.50	0.35	0.15	0.50	0.65	0.15
0.60	0.44	0.16	0.60	0.74	0.14
0.70	0.55	0.15	0.70	0.82	0.12
0.80	0.68	0.12	0.80	0.88	0.08
0.90	0.83	0.07	0.90	0.94	0.04

4. Conclusions

In international assessments, problems related to translation, differences in curriculum, or incomparability of grade or age levels cannot perhaps be completely overcome. However, the validity of comparisons can be increased through the construction of more comparable tests by eliminating translation-related problems. This study explored the use of DIF methods to detect items that may cause differential response patterns between compared groups because of translation. The findings indicate that DIF methods can be used to identify items that lose their original meaning as a result of translation. Three problems related to translation as identified here are: (i) Differential frequency, difficulty or commonness of vocabulary; (ii) Differential length or complexity of sentences; and (iii) Differential contextual meaning of vocabulary. If bilingual individuals were to participate in the item writing process or if items were screened for such possible pitfalls, some of these problems could be alleviated. Additionally, these DIF items violated some of the rules developed by Brislin et al. (1973) for translatable English. These problems could perhaps also have been avoided if these rules had been applied.

DIF due to translation, cultural, or curriculum-related differences in an item means that the item is not assessing what it is intended to assess. Therefore, these differences adversely affect the accuracy and the validity of the test. The results of both IEA and IAEP studies showed that there was stronger DIF in favor of the groups who took tests in the original languages of the tests. In addition, the internal consistency coefficient was higher for the English-speaking group. Overall fairness and accuracy of tests might be improved if some of the items originated in the native languages of the countries being tested rather than solely in English.

In addition to the above precautions to avoid translation-related problems, DIF analyses can be useful at the pilot stage of the test development, or even after the test has been administered. At the pilot stage, the items that are statistically identified as showing DIF can either be revised or eliminated. After the test has been administered,

DIF methods can be used to identify problematic items that might need to be excluded from international comparisons.

Acknowledgements

I would like to thank Edward Haertel, Wendy Yen and Rebecca Zwick for their valuable suggestions to the previous versions of this paper.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, *34*, 807–816.

Brislin, R. W. (1988). The wording and translation of research instruments. In W. Lonner, & J. W. Berry (Eds.), *Field methods in cross-cultural research.*

Brislin, R. W., Lonner, W., & Thorndike, R. M. (1973). *Cross-cultural research methods*., New York: Wiley.

Cabello, B. (1983). A description of analysis for the identification of potential sources of bias in dual language achievement tests. *The Journal for the National Association for Bilingual Education*, *7*, 35–52.

ETS (1988). *A world of differences, An international assessment of mathematics and science*. Princeton, NJ: Author.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the Military Testing Association.*

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Homewood, IL: Dow Jones/Irwin.

Rodrigues-Bou, I. (1950). *A study of the parallelism of English and Spanish vocabularies*. Puerto Rico: Superior Educational Council of Puerto Rico.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History assessment. *Journal of Educational Measurement*, *26*, 55–66.

Kadriye Ercikan is an Assistant Professor of Measurement, Evaluation and Research Methods, Faculty of Education, University of British Columbia. She specializes in item response theory, large-scale assessments, evaluation methods and translation effects in international assessments. Her publications in the area of measurement have appeared in major journals and she has delivered presentations at numerous educational conferences. She is a member of the NCME International Testing Committee.