Translation procedures in OECD/PISA 2000 international assessment

Aletta Grisay *former senior researcher at University of Liège, Belgium and consultant to the PISA consortium*

As a part of the OECD Programme for International Student Assessment (PISA), an international survey was conducted in 2000 to assess the Reading, Mathematics and Science literacy of 15-year-old students in 31 countries (see McQueen and Mendelovits, this issue; see Grisay, 2002). The article describes the procedures implemented by the PISA International Co-ordination Centre for the development of national versions of the assessment instruments in all instruction languages used in the participating countries. It also presents data (collected during the field trial of the instruments) that provide some empirical information on the effectiveness of these procedures. The International Centre developed two source versions (in English and French) of the instruments. It was recommended that the national adaptation teams produce two independent translations (one from the English and the other from the French source version) of the assessment material into the language of instruction in their country and that they reconcile them into a single national version. A group of international verifiers appointed and trained by the International Centre then checked the equivalence of all national versions against the source versions.

I Introduction

Translation errors are known to be a major cause for items to function poorly in international tests. They are much more frequent than other problems, such as clearly identified discrepancies due to cultural biases or curricular differences.

If a survey is done merely to rank students or countries, this problem can usually be dealt with by developing and pre-testing a larger pool of test items than needed for the assessment. Flawed or unstable items are identified and dropped on the basis of the field trial statistics. The item pool used in the main study would still contain sufficient material, and even if a few residual flaws require new deletions at that stage, they will be unlikely to affect the overall estimate of a country's mean in any significant way.

However, in surveys like the OECD Programme for International

Address for correspondence: Aletta Grisay, 22, avenue des Canadiens, F-94410 Saint-Maurice, France; email: agrisay@attglobal.net

Language Testing 2003 20 (2) 225–240 10.1191/0265532203lt254oa © 2003 Arnold

Student Assessment (PISA), where the aim is to develop descriptive scales, enlarging the initial item pool may not be a sufficient solution, due to the additional requirement that the surviving items adequately cover all framework cells and all described levels in the scales. In this case translation errors are of greater concern, since unstable item characteristics can severely bias the interpretation of the scales and their comparability across countries. Strict verification procedures for translation equivalence therefore need to be implemented.

This article describes the quality assurance procedures that were used in the PISA 2000 assessment to ensure equivalence in 34 national versions of test materials developed by 31 countries in 20 different languages. The procedures included, in particular, the development of two parallel source versions (in English and French), with a recommendation that each country should develop two independent versions in their instruction language (one from each source language), then reconcile them into one national version. Both source versions included systematic information aimed at clarifying the intent, scope and characteristics of each test item, and frequent translation notes for possible translation or adaptation problems.

A document describing the recommended translation procedure and containing detailed translation/adaptation guidelines was provided to participating countries, and used as instruction material in a training session attended by key staff from each national translation team. All national versions were then submitted for central verification against the source versions to a group of international verifiers appointed and specially trained by the PISA International Project Centre. This verification team comprised professional translators proficient in both English and French, with native command of the target language used in the national version submitted to each of them. After entering the corrections proposed by their verifier (or sometimes discussing and rejecting a few of them, or finding alternative solutions), the participating countries were asked to return hard copies of their future test booklets, so that the verifier could perform a final check on accuracy of edits, correct assembly of the material, layout and rendering of graphics.

All participating countries were asked to establish a National Expert Committee, who were in charge of reviewing the appropriateness of the source material for the country's 15-year-old students, helping national translators with terminology and other content-specific problems, as well as reviewing and endorsing the final national version.

II Why double translation from two source languages?

A back translation procedure is the most frequently used to ensure linguistic equivalence of test instruments in international surveys. It requires translating the source version of the test (generally in English) into versions for the national languages, then translating them back to, and comparing them with, the source language to identify possible discrepancies.

This technique is relatively effective for detecting mistranslation or major interpretation problems (Hambleton, 1994; 2002). For example, in the original English version of one of the PISA reading texts proposed for the field trial, no lexical or grammatical clue enabled the reader (and the translator) to identify the main character's gender. Many languages, however, have unambiguous morphological markers that impose a gender choice in almost every sentence in which the character occurs. The comparison of several back translations almost inevitably brings out this type of problem.

Back translation has a serious deficiency, however, which has often been pointed out. In many cases, a translated passage is incorrect because it is too literally transposed, but there is a fairly high risk that the back translation will merely recover the original text without revealing the error. An interesting example is a passage from Somerset Maugham's short story *The Ant and the Grasshopper*, which was proposed as a reading text for the PISA 2000 field trial. Both translators who worked on the development of the French source version translated the content of the italicized sentence in the following passage word for word:

> In this admirable fable (*I apologise for telling something which everyone is politely, but inexactly, supposed to know*) the ant spends a laborious summer gathering its winter store, while the grasshopper sits on a blade of grass singing to the sun.

Translation 1:

> Dans cette fable remarquable (que le lecteur me pardonne si je raconte quelque chose que chacun est courtoisement censé savoir, mais pas exactement), la fourmi consacre un été industrieux à rassembler des provisions pour l'hiver, tandis que la cigale le passe sur quelque brin d'herbe, à chanter au soleil.

Translation 2:

> Dans cette fable admirable (veuillez m'excuser de rappeler quelque chose que chacun est supposé connaître par politesse mais pas précisément), la fourmi passe un été laborieux à constituer des réserves pour l'hiver, tandis que la cigale s'installe sur l'herbe et chante au soleil.

Both translations are literally correct, and would back-translate into an English sentence quite parallel to the original sentence. However, both are semantically unacceptable, since neither reflects the irony of 'politely, but inexactly supposed to know'. The two versions were reconciled and eventually translated into:

> Dans cette fable remarquable (que l'on me pardonne si je raconte quelque chose que chacun est supposé connaître – supposition qui relève de la courtoisie plus que de l'exactitude), la fourmi consacre un été industrieux à rassembler des provisions pour l'hiver, tandis que la cigale le passe sur quelque brindille, à chanter au soleil.

Both translations 1 and 2 would probably appear to be correct when back-translated, whereas the reconciled version would appear further from the original. It is also interesting to note that both French translators and the French reconciler preferred (rightly so, for a French-speaking audience) to revert to the *cigale* [i.e., *cicada*] of La Fontaine's original text, while Somerset Maugham adapted the fable to his English-speaking audience by referring to a *grasshopper* [which would have been *sauterelle* in French]. However, both translators neglected to draw the entomological consequences from their return to the original: they were too faithful to the English text and allowed a strictly arboreal insect [the cicada] to live on a *brin d'herbe* [i.e., a *blade of grass*], that the reconciler replaced by a *brindille* [i.e., a *twig*]. In such a case, the back translation procedure would consider *cigale* and *brindille* as deviations or errors.

A double translation procedure (i.e. two independent translations from the source language, and reconciliation by a third person) offers two significant advantages in comparison with the back translation procedure. First, equivalence of the source and target languages is obtained by using three different people (two translators and one reconciler) who all work on the both the source and the target versions. In the back translation procedure, by contrast, the first translator is the only one who focuses simultaneously on the source and target versions (Hambleton and Kanjee, 1995). Secondly, possible discrepancies are recorded directly in the target language instead of in the source language, as would be the case in a back translation procedure. The examples above are deliberately borderline cases, where both translators happened to make a common error (that could perhaps have been overlooked, had the reconciliation have been less accurate). But the probability of detecting errors is obviously considerably higher when three people rather than one compare the source language with the target language.

PISA used double translation from two different languages because both back translation and double translation procedures come short in that the equivalence of the various national versions depends exclusively on their consistency with a single source version (in general, English). This leads to implicitly giving more weight than would be desirable to those cultural forms related to the reference language. Furthermore, one would wish for as purely a semantic equivalence as possible (since the principle is to measure access that students from different countries would have to the same meaning, through written material presented in different languages). However, using a single reference language is likely to give more importance than would be desirable to the formal characteristics of that language. If a single source language is used, its lexical and syntactic features, stylistic conventions and typical organizational patterns of ideas within the sentence will have more impact than desirable on the target language versions.

Another expected benefit from using two source languages is that it helps in monitoring the degree of freedom to take with respect to a source text. A translation that is too faithful may appear awkward; if it is too free or too literary, it is very likely to fail to be equivalent. Having two source versions in different languages (for which the translation fidelity/freedom has been carefully calibrated and approved by PISA international experts) provides benchmarks for a national reconciler that are far more accurate in this respect, something that neither back translation nor double translation from a single language could provide. In addition, many translation problems are due to idiosyncrasies: words, idioms or syntactic structures in one language appear untranslatable into a target language. The opportunity to consult a second source version can often provide hints at solutions.

Similarly, resorting to two different languages will, to a certain extent, tone down problems linked to the impact of cultural characteristics of a single source language. Admittedly, both languages used here share an Indo-European origin, which may be regrettable in this particular case. However, they do represent sets of relatively different cultural traditions, and are both spoken in several countries with different geographic locations, traditions, social structures and cultures.

Nevertheless, as all major international surveys prior to PISA always used English as their source language, empirical evidence on the consequences of using an alternative reference language was lacking. As far as we know, the only interesting findings in this respect were reported in the IEA/Reading Comprehension survey (Thorndike, 1973), which showed better item coherence (factorial structure of the tests, distribution of the discrimination coefficients) between English-speaking countries than across other participating countries.

From this perspective, using two source languages in PISA was an experimental procedure, which raised two important issues:

- To what extent can two sufficiently 'equivalent' English and

French source versions be developed for use by national translators? Equivalence is a crucial requirement, since any serious discrepancy between the source versions would be likely to affect the quality of all other versions. In addition, possible differences in the overall difficulty of the English and French versions might undermine the general comparability of the instruments derived from them.

- Does the recommended procedure in fact result in better national versions than other possible procedures, which might be less costly and/or time-consuming for both the national project teams and the International Co-ordination Centre?

The analyses presented below used the PISA/2000 field test data to provide information on these two issues.

Since Reading was the major domain in the PISA/2000 assessment, it was particularly crucial that the stimuli used in the test units be as equivalent as possible in terms of their linguistic difficulty. We compared the length, and a few other characteristics, of the English and French versions of a sample of PISA stimuli, using readability formulas to assess their relative difficulty in the two languages.

The field trial data from the English-speaking and French-speaking countries was used to check whether the psychometric characteristics of the items in the versions adapted from the English source were similar to those in the versions adapted from the French source. In particular, we wanted to know whether any items showed flaws in all or most of the French-speaking countries, but none in the English-speaking countries, or vice versa.

Also based on the field trial statistics, a comparison was done between the national versions developed through the recommended procedure and those obtained through alternative procedures, in order to identify the translation methods that produced less flawed items than others did.

III Linguistic characteristics of the English and French source versions of the stimuli in the PISA material

Most of the text passages used as stimuli in the PISA field trial, and part of the items, were contributed by the participating countries: the retained material contained submissions from 18 countries. The material submitted for inclusion had to be either in English or in French. However, less than 5% of the submissions received were in French. As a consequence, the development of an alternative source version was mainly done through double-translating and reconciling into French an essentially English ‘first’ source version of the material.

1 Differences in length

The length of stimuli was compared using the texts included in 50 reading and 7 science test units. No mathematics units were included, since these units had virtually no text, or very short texts. Some of the reading and science units were also excluded, since they had texts that were too short or just tables.

The French version of the stimuli proved to be significantly longer than the English version. On average, the number of words in the French stimuli was 410, as compared to 367 in the English stimuli, that is, 12% more words. In addition, the average length of words is higher in French (5.09 characters per word, vs. 4.83 characters per word in English). As a result, the total character count was considerably higher in French (on average, 18.84% more characters in the French than in the English version of the sampled stimuli).

This did not affect the relative length of the passages in the two languages: the correlation between the English (ENG) and French (FRE) word count was 0.99, as was that between the ENG and FRE character count.

However, some variation was observed from text to text in the 'increase' in length from English to French. There was some evidence that those texts that were originally in French or in languages other than English tended either to have fewer words in the French version than in English (*Pole Sud*: –2%; *Shining Object*: –4%; *Macondo*: – 5%) or to show only minor differences (*Police*: +2.5%; *Amanda and the Duchess*: +1.2%; *Rhinoceros*: +5%; *Just Judge*: +1.6%; *Corn*: +4%).

2 Effects of differences in text length on the difficulty of the test items

Forty-nine prose texts of sufficient length (more than 150 words in the ENG version) were retained for an analysis on the possible effects of the higher word count associated with the translation into French. For 10 of them, the FRE version was 'shorter' than the ENG version, or very similar in length (less than 5% increase in the word count). Ten others had a 6–10% increase in the FRE word count. Fifteen had an increase from 11–20%, and the remaining 14 had an increase of more than 20%.

Eleven PISA countries had English or French as (one of) their instruction language(s). The ENG source version of the field trial instruments was used, with a few national adaptations, in Australia, Canada (English), Ireland, New Zealand, the UK and the USA. The French source version was used, with a few national adaptations, in

Belgium (French community), Canada (French), France, Luxembourg and Switzerland (French).

However, the booklets used in Luxembourg contained both French and German material, which made it impossible to use the statistics from Luxembourg for a comparison between the ENG and FRE data. Therefore the analysis was done using the item statistics from the remaining 10 countries: six countries in the 'Adaptation from English' group, and four countries in the 'Adaptation from French' group.

In the field trial, the students answered a total of 277 questions related to the selected stimuli (5 or 6 questions per text). The overall percentage of correct answers to the subset of questions related to each text was computed for each country in each language group. Table 1 shows the average results by type of stimuli. In both ENG and FRE countries the item difficulty appeared to be higher in the group of test units where the FRE stimuli presented only a minor increase in length compared to the ENG version, while the groups with significant increase in length proved to be easier.

The mean percentage of correct answers was slightly higher in the ENG countries for all groups of test units, but more so for the groups of units containing the stimuli that had the largest increase in the FRE version. This 'group by language' interaction was significant ($F = 3.62$, $p < .05$), indicating that the 'longer' FRE units tended to be more difficult for the French-speaking students than those with only a minor increase in word count.

This pattern of results suggests that the burden added to the reading

Table 1 Percentage of correct answers in ENG and FRE countries for groups of test units with small or large differences in the length of stimuli in the source languages

	ENG countries		FRE countries		All	
Increase in word count in FRE, compared to the ENG version	Mean percentage of correct answers	S.D.	Mean percentage of correct answers	S.D.	Mean percentage of correct answers	S.D.
a) Less than 5% increase in FRE (10 units)	59.7%	11.3	58.1%	13.4	58.9%	12.1
b) Between 6 and 10% (10 units)	63.1%	13.4	59.1%	15.0	61.1%	14.0
c) Between 11 and 20% (15 units)	65.0%	13.7	62.1%	14.1	63.5%	13.7
d) More than 20% (14 units)	68.2%	11.8	63.7%	12.6	66.0%	12.2

Note: $n = 490$; 10 countries × 49 tests tasks in countries using the 'longer' version may well have had some (modest) effect on the students' performance.

3 Linguistic difficulty

Readability indices were computed for a subsample of 26 texts, both in the French version (using the Flesch-De Landsheere and Henry formulas; see Flesch, 1948; De Landsheere, 1973; Henry, 1975) and in the English version (using the Fry, Dale and Chall and Spache formulas; see Dale and Chall, 1948; Spache, 1972; Fry, 1980). Most of these formulas use similar indicators to quantify the linguistic 'density' of the texts. The most common indicators, included in both the French and English formulas, are:

- average length of words: an indicator of lexical difficulty;
- percentage of low-frequency words: mostly an indicator of abstractness; and
- average length of sentences: an indicator of syntactic complexity.

However, the metrics are not the same across languages, and a number of other ingredients – sometimes language specific – are used in each of the formulas, which prevents true direct comparison of the indices obtained for the English and French versions of the same text. Therefore, the means in Table 2 must be considered with caution, while the correlations between the ENG and FRE indices are more reliable.

All of the correlations were reasonably high or very high, indicating that the stimuli that had higher indices of linguistic difficulty in ENG also tended to have higher difficulty indices than other passages in FRE. That is, ENG texts with more abstract or more technical vocabulary, or with longer and more complex sentences, etc., tended to show the same characteristics when translated into French.

IV Psychometric quality of the French versions of the material, compared to the English versions

Using the statistics from the field trial item analyses, all items containing one or more of the following flaws were identified in each national version of the instruments:

- items with DIF, i.e., significantly easier (or harder) than in most other versions;
- items with too large a fit (greater than 1.20);
- items with too low a discrimination index (less than 0.15).

Some 30 items (out of the 561 reading, mathematics and science

Table 2 Correlation of linguistic difficulty indicators in 26 ENG and FRE texts

	ENG		FRE		r (ENG/FRE)
	Mean	S.D	Mean	S.D	
Average word length	4.83 characters	0.36	5.09 characters	0.30	0.60
Percentage of low-frequency words	18.7%	8.6	21.5%	5.4	0.61
Average sentence length	18 words	6.7	21 words	7.1	0.92
Average readability index	Dale–Chall: 32.84	Dale–Chall: 9.76	Henry: 0.496 Flesch–De Landsheere: 38.62	Henry: 0.054 Flesch–De Landsheere: 14.18	0.72 0.83

items included in the field trial material) appeared to have problems in all 31 participating countries, or in most of them. Since in these cases one can be rather confident that the flaw had to do with the content of the item – not with the quality of the translation – all observations related to these items were discarded from the comparisons. In addition, a few items that had no statistics in some of the countries (items with 0% or 100% correct answers) had to be discarded. Table 3 shows the distribution of flawed items in the English-speaking and French-speaking countries.

As Table 3 clearly shows, the pattern was very similar in the two groups of countries. The percentage of flawed items in the field trial material varied from 5.8% to 9.4% in the ENG countries, and from 5.3% to 8.9% in the FRE countries, with almost identical means (ENG: 7.5%, FRE 7.7%; $F = 0.05$, $p < 0.83$).

The detailed item statistics show that only one reading item (R228Q03) was flawed in all four FRE countries, but in none of the ENG countries. Three other items (R069Q03A, R069Q05 and R247Q01) were flawed in three out of the four FRE countries, but in none of the ENG countries. Conversely, only four items had flaws in all of the six ENG countries or in four or five of them, but only in one of the FRE countries (R070Q06, R085Q06, R088Q03, R119Q10). None of the science or mathematics items showed the same kind of imbalance.

Table 3 Percentage of flawed items in the ENG and FRE national versions

	Number of items	Too easy	Too hard	Large fit	Low discrimination	Percentage of items with bugs
ENG countries						
Country A	532	0.4	0.0	5.6	3.0	7.7
Country B	532	0.0	0.6	3.4	3.6	6.6
Country C	530	0.0	0.0	8.9	1.5	9.4
Country D	527	0.0	0.4	5.9	3.6	7.8
Country E	532	0.2	0.0	6.8	1.9	7.5
Country F	532	0.4	0.6	4.1	1.7	5.8
Mean ENG countries	3184	0.1	0.3	5.8	2.5	7.5
FRE countries						
Country G	531	0.0	0.6	5.6	3.8	7.9
Country H	531	0.0	0.2	5.6	5.6	8.9
Country I	532	0.2	0.9	2.8	2.4	5.3
Country J	530	0.0	0.2	3.6	6.2	8.7
Mean FRE countries	2124	0.04	0.4	4.4	4.5	7.7
ALL	5309	0.1	0.3	5.2	3.3	7.6

V Psychometric quality of the national versions obtained through the recommended or through alternative procedures

In countries that had instruction languages other than ENG and FRE, the national versions of the PISA field trial instruments were developed either through the recommended procedures, or through one of a number of alternative methods:

- Double translation from ENG and FRE: This was the recommended procedure. It was fully implemented (that is, for all three domains and for both the stimuli and the items) in Norway, Iceland, Sweden, Hungary, the Netherlands and Belgium (Flemish).
- Double translation from ENG, with cross-checks against FRE: This was done in Denmark, Finland, Poland and in the German-speaking countries (Austria, Germany, Switzerland (German)).
- Double translation from ENG, without cross-checks against FRE: this was done in the Spanish and Portuguese countries (Spain, Mexico, Portugal and Brazil).
- Single translation from ENG or from FRE: this was done in Greece, Korea, Latvia and Russia. Most of the material was single-translated from ENG in these countries. In Greece and in Latvia, a small part of the material was single-translated from FRE and the rest from ENG.
- Mixed methods, e.g., Luxembourg had bilingual booklets, with the FRE material adapted from FRE and the GER material adapted from the 'common' German version used by all German-speaking countries. Japan had the reading stimuli double-translated from ENG and FRE and the reading items as well as the mathematics and science material double-translated from ENG; Italy and Switzerland (Italian) single-translated the material, one from ENG, the other from FRE, with a view to reconciling the two versions, but they ran out of time and were able to reconcile only part of the reading material. Therefore they both kept the remaining units single-translated, with some checks against the version derived from the other source language.

In each country, the items with flaws were identified, using the same procedure as for the ENG and FRE countries, in order to find out whether some of these methods resulted in better national versions than others (i.e., with a smaller proportion of flawed items). Table 4 shows the percentage of flawed items observed by method and by country. As Table 4 shows, there was significant between-country variation in the number of flawed items in each group, which indicates that the method used was by no means the unique determinant of the psychometric quality achieved in the development of the instruments.

Table 4 Percentage of flawed items by translation method

Method	Mean percentage of items with flaws	Country	*n*	Percentage of items with flaws
a) Adaptation from source version	7.6	Adapted from ENG	3185	7.5
		Adapted from FRE	2124	7.7
b) Double translation from ENG and FRE	8.0	Country 1	532	9.0
		Country 2	532	7.5
		Country 3	532	7.3
		Country 4	532	10.3
		Country 5	532	7.0
		Country 6	532	7.0
c) Double translation from ENG (with cross-checks against FRE)	8.8	Country 7	532	5.6
		Country 8	532	10.5
		Country 9	532	7.9
		Country 10	532	7.5
		Country 11	532	8.5
		Country 12	532	12.8
f) Double translation from ENG (without use of FRE)	12.1	Country 13	532	13.9
		Country 14	532	8.3
		Country 15	532	16.0
		Country 16	532	10.3
d) Single Translation	11.1	Country 17	532	9.4
		Country 18	532	16.0
		Country 19	532	9.2
		Country 20	532	9.8
e) Other (mixed methods)	10.3	Country 21	532	13.9
		Country 22	532	9.4
		Country 23	532	9.2
		Country 23	532	13.2
		Country 24	532	5.6

Most probably, other important factors were the accuracy of the national translators and reconcilers, as well as the quality of the work done by the international verifiers.

However, the data seem to confirm the hypothesis that the recommended procedure (Method b: double translation from ENG and FRE; see Table 5) produced national versions that did not differ significantly from the versions derived through adaptation from one of

Table 5 Differences between translation methods

	b) Double translation English and French	c) Double translation with checks	d) Other methods	e) Single translation	f) Double translation with no checks
a) Adapted from sources	A > B $F = 0.45$ $p = 0.51$	A > C $F = 1.73$ $p = 0.21$	A > D $F = 5.23$ $p = 0.04$	A > E $F = 4.24$ $p = 0.06$	A > F $F = 13.79$ $p = 0.02$
b) Double translation English and French		B > C $F = 0.45$ $p = 0.52$	B > D $F = 2.27$ $p = 0.17$	B > E $F = 4.37$ $p = 0.07$	B > F $F = 7.12$ $p = 0.03$
c) Double translation and checks			C > D $F = 0.69$ $p = 0.43$	C > E $F = 1.58$ $p = 0.24$	C > F $F = 3.14$ $p = 0.11$
d) Other methods				D > E $F = 0.14$ $p = 0.72$	D > F $F = 0.66$ $p = 0.44$
e) Single translation					E > F $F = 0.19$ $p = 0.68$

the source languages in terms of the number of flaws. Double translation from only one language appeared also to be effective when it was accompanied by extensive cross-checks against the other source (Method c).

The average number of flawed items was higher in all other groups of countries than in those that used both sources (either by double translating from the two languages or by using one of the sources for double translation and the other for cross-checks). Method e (single translation) and method f (Double translation from only one language, without cross-checks) proved to be the least effective methods (Table 5).

VI Discussion

The analyses above show that the relative linguistic complexity of the reading stimuli in the English and French field trial versions of the PISA assessment materials (as measured using readability formulas) was reasonably comparable. However the absolute differences in word and character counts between the two versions were significant, which had a (modest) effect on the difficulty of the items associated with those stimuli that were much longer in the French version than in the English version.

The average length of words and sentences is a characteristic that differs across languages and that, obviously, cannot be entirely controlled by translators when they adapt test instruments. In this respect, it is probably an impossible task to develop 'true' equivalent versions of tests that involve large amounts of written material. However, languages that are 'longer' than others often are so because they have slightly more redundant morphological or syntactic characteristics, which may help compensate for part of the burden added on reading tasks, especially in test situations like PISA, that had no strong speededness requirements.

No significant differences were observed between the two source versions in the overall number and distribution of flawed items. With a few exceptions, the number of translation flaws in the field trial national versions developed by the participating countries through translation from the source materials remained acceptable: only 9 countries had more than 10% flawed items, compared to the average 7.6% observed in countries that used one of the source versions with just small national adaptations.

The data seemed to support the hypothesis that using double translation from both source versions would result in better translations, with a lesser incidence of flaws than when using only one of the sources. An alternative procedure which also appeared to be effective was double translation from one source with extensive cross-checks against the other source.

VII References

Dale, E. and **Chall, J.S.** 1948: *A formula for predicting readability*. Columbus, OH: Bureau of Educational Research, Ohio State University.

De Landsheere, G. 1973: *Le test de closure, mesure de la lisibilité et de la compréhension.* Bruxelles: Nathan-Labor.

Flesch, R. 1948: A new readability yardstick. *Journal of Applied Psychology* 32, 221–33.

Fry, E. 1980: Graph for estimating readability. *Canadian Library Journal* 37, 249.

Grisay, A. 2002: Translation and cultural appropriateness of the test and survey material. In Adams, R. and Wu, M., editors, *PISA 2000 Technical Report*. Paris: Organisation for Economic Co-operation and Development.

Hambleton, R.K. 1994: Guidelines for adapting educational and psychological tests: a progress report. *European Journal of Psychological Assessment* 10, 229–44

Hambleton, R.K. 2002: Adapting achievement tests into multiple languages for international assessments. In Porter, A.C. and Gamoran, A., editors, *Methodological advances in cross national surveys of educational achievement*. Washington DC: National Academy Press.

Hambleton, R.K. and **Kanjee, A.** 1995: Translating tests and attitude scales. In Husen, T. and Postlethwaite, T.N., editors, *International encyclopedia of education*. 2nd edition. Oxford: Pergamon Press, 6328–334.

Henry, G. 1975: *Comment mesurer la lisibilité.* Bruxelles: Labor.

OECD/PISA 1999: Translation of test instruments and survey material, in *PISA/2000 field trial national project manager's manual*. Camberwell: Australian Council for Educational Research, 21–54.

—— 2000: Translation of test instruments and survey material, in *PISA/2000 main test national project manager's manual.* Camberwell: Australian Council for Educational Research, 31–49.

Spache, G.D. 1972: *Good reading for poor readers.* 8th edition. Champaign, IL: Garrard Publishing.

Thorndike, R.L. 1973: *Reading comprehension education in fifteen countries*. Uppsala: Almqvist and Wiksell.