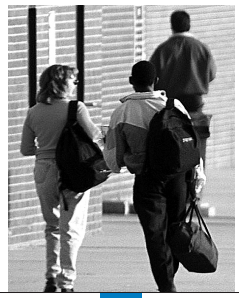

5

Translation and cultural appropriateness of the test and survey material

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

INTRODUCTION

Literature on empirical comparative research refers to translation issues as one of the most frequent problems in cross-cultural surveys. Translation errors are much more frequent than other problems, such as clearly identified discrepancies due to cultural biases or curricular differences. (Harkness, Van de Vijver and Mohler, 2003; Hambleton, Merenda and Spielberger, 2005).

If a survey is done merely to rank countries or students, this problem can be avoided somewhat since once the most unstable items have been identified and dropped, the few remaining problematic items are unlikely to affect the overall estimate of a country's mean in any significant way.

The aim of PISA, however, is to develop descriptive scales, and in this case translation errors are of greater concern. The interpretation of a scale can be severely biased by unstable item characteristics from one country to another. One of the important responsibilities of PISA is therefore to ensure that the instruments used in all participating countries to assess their students' literacy provide reliable and fully comparable information. In order to achieve this, PISA implemented strict verification procedures for translation/adaptation and verification procedures.

These procedures included:

- Development of two source versions of the instruments (in English and French);
- Double translation design;
- Preparation of detailed instructions for the translation of the instruments for the field trial and for their review for the main study;
- Preparation of translation/adaptation guidelines;
- Training of national staff in charge of the translation/adaptation of the instruments;
- Verification of the national versions by international verifiers.

DEVELOPMENT OF SOURCE VERSIONS

Part of the new test materials used in PISA 2006 was prepared by the consortium test development teams on the basis of the submissions received from the participating countries. Items were submitted by 21 different countries, either in their national language or in English. The other part of the material was prepared by the test development teams themselves in CITO, NIER, ILS, IPN and ACER. Then, all materials were circulated (in English) for comments and feedbacks to the Expert Groups and the NPMs.

The item development teams received specific information/training about how to anticipate potential translation and cultural issues. The document prepared for that purpose was mainly based on experience gained during previous PISA cycles. The items developers used it as reference when developing and reviewing the items.

The French version was developed at this early stage through double translation and reconciliation of the English materials into French, so that any comments from the translation team could, along with the comments received from the Expert Groups and the NPMs, be used in the finalisation of both source versions.

Experience has shown that some translation issues do not become apparent until there is an attempt to translate the instruments. As in previous PISA cycles, the translation process proved to be very effective

in detecting residual errors overlooked by the test developers, and in anticipating potential translation problems. In particular, a number of ambiguities or pitfall expressions could be spotted and avoided from the beginning by slightly modifying both the English and French source versions; the list of aspects requiring national adaptations could be refined; and further translation notes could be added as needed. In this respect, the development of the French source version served as a pilot translation, and contributed to providing National Project Managers with source material that was somewhat easier to translate or contained fewer potential translation problems than it would have had if only one source had been developed.

The final French source version was reviewed by a French domain expert, for appropriateness of the science terminology, and by a native professional French proof-reader for linguistic correctness. In addition, an independent verification of the equivalence between the final English and French versions was performed by a senior staff member of cApStAn who is bilingual (English/French) and has expertise in the international verification of the PISA materials, and used the same procedures and verification checklists as for the verification of all other national versions.

Finally, analyses of possible systematic translation errors in all or most of the national versions adapted from the French source version were conducted, using the main study item statistics from the five French-speaking countries participating in PISA 2006.

DOUBLE TRANSLATION FROM TWO SOURCE LANGUAGES

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

A double translation design (*i.e.* two independent translations from the source language(s), and reconciliation by a third person) offers two significant advantages in comparison with the back translation design:

- Equivalence of the source and target versions is obtained by using three different people (two translators and a reconciler) who all work on both the source and the target versions. In a back translation design, by contrast, the first translator is the only one to simultaneously use the source and target versions;
- Discrepancies are recorded directly in the target language instead of in the source language, as would be the case in a back translation design.

PISA uses double translation from two different languages because both back translation and double translation designs fall short in that the equivalence of the various national versions depends exclusively on their consistency with a single source version (in general, English). In particular, one would wish the highest possible semantic equivalence (since the principle is to measure access that students from different countries would have to a same meaning, through written material presented in different languages). However, using a single reference language is likely to give undue importance to the formal characteristics of that language. If a single source language is used, its lexical and syntactic features, stylistic conventions and the typical patterns it uses to organise ideas within the sentence will have a greater impact on the target language versions than desirable (Grisay, 2003).

Some interesting findings in this respect were reported in the IEA/reading comprehension survey (Thorndike, 1973), which showed a better item coherence (factorial structure of the tests, distribution of the discrimination coefficients) between English-speaking countries than across other participating countries.

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

Other anticipated advantages of using two source languages in the PISA assessment included:

- Many translation problems are due to idiosyncrasies: words, idioms, or syntactic structures in one language appear untranslatable into a target language. In many cases, the opportunity to consult the other source version may provide hints at solutions;
- The desirable or acceptable degree of translation freedom is very difficult to determine. A translation that is too faithful may appear awkward; if it is too free or too literary it is very likely to jeopardise equivalence. Having two source versions in different languages (for which the translation fidelity/freedom has been carefully calibrated and approved by consortium experts) provides national reconcilers with accurate benchmarks in this respect, and that neither back translation nor double translation from a single language could provide.

Since PISA was the first major international survey using two different source languages, empirical evidence from the PISA 2000 field trial results was collected to explore the consequences of using alternative reference languages in the development phase of the various national versions of the survey materials. The outcomes of this study were reported in Chapter 5 of the *PISA 2000 Technical Report* (Adams and Wu, 2002; Grisay, 2003).

PISA 2003 main study data analyses were used to identify all items showing even minor weaknesses in the seven English-speaking countries or communities and the five French-speaking countries or communities that developed their national versions by just entering national adaptations in one of the source versions provided by the consortium (OECD 2005). Out of the 167 items used in the main study, 103 had no problems in any of the French and English versions and 29 had just occasional problems in one or two of the twelve countries. Thirteen items had weak statistics in both English and French versions but also appeared to have flaws in at least half of the participating countries. No items had weaknesses in all French versions and no flaws in any of the English versions. Some imbalance was observed for nine items. In fact the overall percentage of weak items was very similar in both the group of English testing countries and the group of French testing countries.

Empirical evidence on the quality of the national versions obtained was collected by analysing the proportion of weak items in each national data set, based again on the PISA 2003 main study item analyses, and using the same criteria for identifying weak items as for the source versions.

Among countries that used double translation from just one of the source versions, 12.5% of the items were considered weak, compared to 8.5% in countries that used both source versions in their translations, and 6.5% in countries whose versions were derived directly from either the English or French source version. This seems to indicate that double-translation from only one source language may be less effective than double translation from both languages, confirming a trend already observed in PISA 2000.

Due to these results, a double translation and reconciliation procedure using both source languages was again recommended in PISA 2006.

PISA TRANSLATION AND ADAPTATION GUIDELINES

The *PISA Translation and Adaptation Guidelines* as prepared in previous PISA studies were revised to include more detailed advice on translation and adaptation of science materials, and additional warnings about

common translation errors identified during the verification of the PISA 2003 materials and the development of the French source version. These guidelines were revised with a view to obtaining a document that would be relevant to any PISA cycle. The guidelines included:

- Instructions for national version(s): According to the PISA technical standards, students should be tested in the language of instruction used in their school. Therefore, the NPMs of multilingual countries were requested to develop as many versions of the test instruments as there were languages of instruction used in the schools included in their national sample. Cases of minority languages used in only a very limited number of schools could be discussed with the sampling referee to decide whether such schools could be excluded from the target population without affecting the overall quality of the data collection;
- Instructions on double or single translation: Double-translation was required for the tests, questionnaires and for the optional questionnaires, but not for the manuals and other logistic material;
- Instructions on recruitment and training: It was suggested, in particular, that translated material and national adaptations deemed necessary be submitted for review and approval to a national expert panel composed of domain specialists;
- Description of the PISA translation procedures: It was required that national versions be developed through double translation and reconciliation with the source material. It was recommended that one independent translator would use the English source version and that the second would use the French version. In countries where the NPM had difficulty appointing competent translators from French/English, double translation from English/French only was considered acceptable according the *PISA Technical Standards 5.1 and 5.2*.

Other sections of the *PISA Translation and Adaptations Guidelines* were intended for use by the national translators and reconcilers and included:

- Recommendations to avoid common translation traps. An extensive section giving detailed examples on problems frequently encountered when translating assessment materials, and advice on how to avoid them;
- Instructions on how to adapt the test material to the national context. This listed a variety of rules identifying acceptable/unacceptable national adaptations and including specific notes on translating mathematics and science material;
- Instructions on how to translate and adapt the questionnaires and manuals to the national context;
- The check list used for the verification of PISA material.

After completion of the field trial, an additional section of the Guidelines was circulated to NPMs, as part of their *Main Study NPM Manual*, together with the revised materials to be used in the main study. This section contained instructions on how to revise their national version(s).

TRANSLATION TRAINING SESSION

NPMs received sample materials to use when recruiting national translators and training them at the national level. The NPM meeting held in September 2004 included a session on the field trial translation/adaptation activities in which recommended translation procedures, *PISA Translation and Adaptation Guidelines*, and the verification process were presented in detail.

TESTING LANGUAGES AND TRANSLATION/ADAPTATION PROCEDURES

NPMs had to identify the testing languages according to instructions given in the *Sampling Manual* and to record them in a sampling form for agreement.

Prior to the field trial, NPMs had to fill in a Translation Plan describing the procedures used to develop their national versions and the different processes used for translator/reconciler recruitment and training. Information about a possible national expert committee was also sought. This translation plan was reviewed by the consortium for agreement and in December 2004 the NPMs were asked to either confirm that the information given was accurate or to notify which changes were made.

Countries sharing a testing language were strongly encouraged to develop a common version in which national adaptations would be inserted or, in the case of minority languages, to borrow an existing verified version. There is evidence from PISA 2000 and 2003 that high quality translations and high levels of equivalence in the functioning of items were best achieved in the three groups of countries that shared a common language of instruction (English, French and German) and could develop their national versions by introducing a limited number of national adaptations in the common version. Additionally, having a common version for different countries sharing the same testing language implies that all students instructed in a given language receive booklets that are as similar as possible, which should reduce cross-countries differences due to translation effects.

Table 5.1 lists countries that shared a common version of test items with national adaptations.

Table 5.1
Countries sharing a common version with national adaptations

Language	Countries	Collaboration
Arabic	Jordan and Qatar	Jordan developed a version in which Qatar introduced adaptations (Field trial only).
Chinese (c)	Hong Kong-China, Macao-China and Chinese Taipei	Commonly developed Chinese version: Two single translations produced by 2 countries and reconciliation by the third one
Dutch	Netherlands, Belgium	Belgium (Flemish Community) introduced adaptations in the verified Dutch version
English	Australia, Canada, Hong Kong-China, Ireland, Qatar, New Zealand, Scotland, Sweden, United Kingdom, USA	Adaptations introduced in the English source version
French	Belgium, Canada, France, Luxembourg, Switzerland	Adaptations introduced in the French source version
German	Austria, Belgium, Germany, Italy, Luxembourg, Switzerland	Adaptations introduced in a commonly developed German version
Hungarian	Hungary, Serbia, Slovak Republic, Romania	For their Hungarian versions, Serbia and the Slovak Republic introduced adaptations in the verified version from Hungary
Italian	Italy, Switzerland, Slovenia	Switzerland (Canton Ticino) and Slovenia introduced adaptations in the verified version from Italy
Russian	Russia, Azerbaijan, Estonia, Kyrgyzstan, Latvia, Lithuania	Adaptations introduced in the verified version from Russia or Kyrgyzstan[1]
Polish	Poland, Lithuania	For its Polish version, Lithuania introduced adaptations in the verified version from Poland
Slovene	Slovenia, Italy	Use of Slovene version in Italy
Portuguese	Portugal, Macao-China	Macao-China introduced adaptations in the verified version from Portugal
Spanish	Mexico, Argentina, Uruguay	Argentina and Uruguay introduced adaptations in the verified version from Mexico
Swedish	Sweden, Finland	For its Swedish version, Finland introduced adaptations in the verified version from Sweden

1. Kyrgyzstan first adapted the version from Russia, then in the Main Study, due to time constraints some countries adapted the verified version from Kyrgyzstan.

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Additionally Chile and Colombia collaborated with each providing one translation (one from English and one from French) to the other. This however did not lead to a common version as each country performed the reconciliation separately.

Table 5.2 summarises the translation procedures as described in the country *Translation Plans.*

Table 5.2
PISA 2006 translation/adaptation procedures

Procedures	Number of national versions
Use one of the source versions with national adaptations	15
Use of a commonly developed version with national adaptations	7
Use of a borrowed verified version with or without national adaptations	19
Double translation from both source versions	16
Double translation from English or French source with cross-checks against the other source version	12
Double translation from English source only	15
Alternative procedures	3

A total of 87 national versions of the materials were used in the PISA 2006 main study, in 44 languages, The languages were: Arabic (4 versions), Azeri, Bahasa Indonesian, Basque, Bulgarian, Catalan, Chinese (3 versions), Croatian, Czech, Danish, Dutch (2 versions); Estonian, English (10 versions), Finnish, French (5 versions), Galician, German (6 versions), Greek, Hebrew, Italian (3 versions), Hungarian (3 versions), Icelandic, Irish, Japanese, Korean, Kyrgyz, Latvian, Lithuanian, Norwegian (Bokmål), Norwegian (Nynorsk), Polish (2 versions), Portuguese (3 versions), Romanian, Russian (5 versions), Serb Ekavian variant, Serb Yekavian variant, Slovak, Slovene (2 versions), Spanish (6 versions), Swedish (2 versions), Thai, Turkish, Uzbek and Valencian.

International verification (described in section below) occurred for 78 national versions out of the 87 used in the main study.

International verification was not implemented when:

- A testing language was used for minorities that make less than 5% of the target population as for Irish, Hungarian (Serbia and Romania), Polish (Lithuania), Valencian. In that case the verification is organised at the national level;
- When countries borrowed a version that had been verified at the national level without making any adaptations as for German (Belgium), English (Sweden), Portuguese (Macao-China), Slovene (Italy), Italian (Slovenia).

INTERNATIONAL VERIFICATION OF THE NATIONAL VERSIONS

As in PISA 2003, one of the most important quality control procedures implemented to ensure high quality standards in the translated assessment materials consisted in having an independent team of expert verifiers, appointed and trained by the consortium, verify each national version against the English and French source versions.

Two verification co-ordination centres were established. One was at ACER in Melbourne (for national adaptations used in the English-speaking countries). The second one was at cApStAn, which has been involved in preparing the French source versions of the PISA materials and verifying non-English national versions since PISA 2000.

The consortium undertook international verifications of all national versions in languages used in schools attended by more than 5% of the country's target population. For languages used in schools attended by 5% or less minorities, international-level verification was deemed unnecessary since the impact on the country results would be negligible, and verification of such languages was more feasible at national level.

For a few minority languages, national versions were only developed (and verified) in the main study phase. This was considered acceptable when a national centre had arranged with another PISA country to borrow its main study national version for their minority (*e.g.* adapting the Swedish version from Sweden for Swedish schools in Finland, the Russian version from the Russian Federation for Russian schools in Latvia), or when the minority language was considered to be a variant that differed only slightly from the main national language (*e.g.* Nynorsk in Norway).

English- or French-speaking countries or communities were allowed to only submit national adaptation forms for verification. This was also considered acceptable, since these countries used national versions that were identical to the source version except for the national adaptations.

The main criteria used to recruit translators to lead the verification of the various national versions were that they had:

- Native command of the target language;
- Professional experience as translators from English or French or from both English and French into their target language;
- Sufficient command of the second source language (either English or French) to be able to use it for cross-checks in the verification of the material;
- Familiarity with the main domain assessed (in this case, science);
- A good level of computer literacy;
- As far as possible, experience as teachers and/or higher education degrees in psychology, sociology or education.

As a general rule, the same verifiers were used for homolingual versions (*i.e.* the various national versions from English, French, German, Italian and Dutch-speaking countries or communities). However, the Portuguese language differs significantly from Brazil to Portugal, and the Spanish language is not the same in Spain and in Latin American countries, so independent native translators had to be appointed for those countries.

In a few cases, both in the field trial and the main study verification exercises, the time constraints were too tight for a single person to meet the deadlines, and additional verifiers had to be appointed and trained.

Verifier training sessions were held prior to the verification of both the field trial and the main study materials. Attendees received copies of the PISA information brochure, *Translation Guidelines*, the English and French source versions of the material and a *Verification Check List* developed by the consortium. The training sessions focused on:

- Presenting verifiers with PISA objectives and structure;
- Familiarising them with the material to be verified;
- Reviewing and extensively discussing the *Translation Guidelines* and the *Verification Check List*;

- Conducting hands-on exercises on specially adapted target versions;
- Arranging for schedules and for dispatch logistics;
- Security requirements.

The verification procedures were improved and strengthened in a number of respects in PISA 2006, compared to previous rounds.

VegaSuite

- For the main study phase, cApStAn developed a web-based upload-download platform known as Vegasuite for file exchange and archiving, to facilitate and automate a number of processes as PISA verification grew in size. This development was well received by NPMs and verifiers.

Documentation

- Science textbooks selected and sent by the National Centres of the participating countries were distributed to verifiers. These textbooks, from the grades attended by most 15-year-olds in the respective countries, were used by verifiers as reference works because the NPMs deemed them representative of the level/register of scientific language familiar to 15-year-olds students in their country.

Verification of test units

- As in previous rounds, verifiers entered their suggested edits in MS Word files, using the track changes mode, to facilitate the revision of verified materials by the NPMs (who could directly accept or refuse the edits proposed). But for all issues deemed likely to affect equivalence between source version(s) and target version, verifiers were also instructed to insert a comment in English at the appropriate location in the test adaptation spreadsheet (TAS). This was to formalise the process by which a) the consortium verification referee is informed of such issues and can liaise as needed with the test developers; b) if there is disagreement with the National Centre (NC), a back-and-forth discussion ensues until the issue is resolved; c) key corrections in test materials are pinpointed so that their implementation can be double-checked at final optical check (FOC) phase. In previous verification rounds, this process took place in a less structured way;
- Following the field trial verification, cApStAn analysed the comments made by verifiers in the TAS, leading to a classification using a relatively simple set of categories. The purpose was to reduce variability in the way verifiers document their verification; to make it easier for the consortium referee to judge the nature of an issue and take action as needed; and to provide an instrument to help assess both the initial quality of national versions and the quality of verifiers' output;
- For the main study phase, an innovation in the TAS was that verifiers used a scroll-down menu to categorize issues in one of 8 standardised verification intervention categories: added information, missing information, layout/visual issues, grammar/syntax, consistency, register/wording, adaptation, and mistranslation. a comments column allowed verifiers to explain their intervention with a back-translation or description of the problem;
- For the main study phase, the consortium's FT to MS revisions were listed in the TAS. For such revisions, the drop-down menu in the verifier intervention column was dichotomous: the verifier had the choice between OK (implemented) or NOT OK (overlooked). In case the change was partially implemented, the verifier would select OK (implemented) and comment on the issue in the verifier comment column. This procedure ensured that the verifier would check the correct implementation of every single FT to MS change.

- Another innovation for the main study phase: at the top of each TAS was a list of recurring terms or expressions that occur throughout the test material, such as Circle Yes or No. Verifiers were asked to keep track of across-unit consistency for these expressions and, at the end of the verification of a full set of units, to choose, in the verifier intervention column, from three options in a drop-down menu: "OK"; "Some inconsistencies"; or "Many inconsistencies".

Verification of the booklet shell

- This had not been a separate component in previous rounds. The booklet shell was dispatched together with a booklet adaptation spreadsheet (BAS) and verified following the same procedure as the test units. This proved very helpful for both the NCs' and the verifiers' work organisation, because it resulted in timely verification of sensitive issues. In previous rounds, the booklet shell was often verified on a rush basis when camera-ready instruments were submitted for final optical check (FOC).

Final optical check

- As in previous rounds, test booklets and questionnaire forms were checked page-by-page as regards correct item allocation, layout, page numbering, item numbering, graphic elements, item codes, footers, etc (classic FOC). As in previous rounds, this phase continues to prove essential in spotting residual flaws, some of which could not have been spotted during the item pool verification;

- An innovation in PISA 2006 was the systematic verification of whether key corrections resulting from the first verification phase were duly implemented. All TAS and BAS containing key corrections were thus also returned to each country with recommendations to intervene on any residual key correction that was overlooked or incorrectly implemented. A similarly annotated QAS was also returned in cases where corrections had been flagged by the consortium staff in charge of reviewing questionnaires, thus requesting follow-up at FOC stage. Note that in PISA 2000 and PISA 2003, National Centres were given the final responsibility for all proposed corrections and edits. Although the FOC brief previously included performing random checks to verify whether crucial corrections proposed during Item Pool verification were duly implemented, in practice this was made difficult by the uncertainty on whether the National Centre had accepted, rejected or overlooked corrections made by the verifier. With the systematic verification of key corrections labelled by the consortium, it was possible to have a quantitative and systematic record of implementation of crucial corrections;

Verification of questionnaires and manuals

- As in PISA 2003, NPMs were required to have their questionnaire adaptation spreadsheet (QAS) and manual adaptaton spreadsheet (MAS) approved by consortium staff before submitting them for verification along with their translated questionnaires and manuals;

- The procedure proved to be effective for questionnaires: the instructions to the verifiers were straightforward and the instruments submitted to their scrutiny had already been discussed extensively with consortium staff by the time they had to verify them. Verifiers were instructed to refrain from discussing agreed adaptations unless the back translation into English of the agreed adaptation inadequately conveyed its meaning, in which case the consortium might have unknowingly approved an inappropriate adaptation;

- A significant improvement in PISA 2006 was that the QAS contained entries for all parts of the questionnaires, including notes and instructions to respondents;

- In the case of manuals, verification continued to be challenging in PISA 2006 because of the greater freedom that countries had in adapting these instruments. Following cApStAn's recommendation after the field trial, it was decided to limit the verification of manuals for the main study to a number of key components. The usefulness and effectiveness of this process remains marginal.

Final check of coding guides

- As in PISA 2003, a verification step was added at the main study phase for the coding guides, to check on the correct implementation of late changes in the scoring instructions introduced by the consortium after the NPM coding seminar. Verifiers checked the correct implementation of such edits. These edits had been integrated into the post-FOC TAS of countries for which the verification was over and in the standard TAS of other countries;
- In line with the innovation for PISA 2006 concerning key corrections, the final check of coding guides included a check on the correct implementation of key corrections located in the scoring rubrics, which had been left pending at booklet FOC stage.

Verification outcomes

In previous cycles, the verification reports contained qualitative information about the national versions and illustrative examples of typical errors encountered by the verifiers. In the PISA 2006 main study, the instruments used to document the verification were designed to generate statistics, and some quantitative data is available. The verification statistics by item and by unit yielded information on translation and adaptation difficulties encountered for specific items in specific languages or groups of languages. This type of information, when gathered during the field trial in the next PISA cycle, could be instrumental in revising items for the main study but would also give valuable information on how to avoid such problems in further cycles.

It also makes it possible to detect whether there are items that elicited many verifier interventions in almost all language groups. When this occurs, item developers would be prompted to re-examine the item's reliability or relevance. Similarly, observing the number of adaptations that the countries proposed for some items may give the item developers additional insight into how difficult it is for some countries to make the item suitable for their students. While such adaptations may be discussed with the consortium, it remains likely that extensively adapted items will eventually differ from the source version (*e.g.* in terms for reading difficulty).

As in previous PISA data collections, the verification exercise proved to be an essential mechanism for ensuring quality even though the national versions were generally found to be of high quality in terms of psychometric equivalence. In virtually all versions, the verifiers identified errors that would have seriously affected the functioning of specific items – mistranslations, omissions, loan translations or awkward expressions, incorrect terminology, poor rendering of graphics or layout, errors in numerical data, grammar and spelling errors.

Link material raised a concern again – in a larger than expected number of countries, it proved to be somewhat difficult to retrieve the electronic files containing the final national version of the materials used in the PISA 2003 main study, from which the link items had to be drawn. The verification team performed a litmus check (convergence check on a sample of link units submitted by the countries versus PISA 2003 main study archive) to determine whether the link units submitted were those actually used in the PISA 2003 test booklets. In a number of cases, the verification team or the consortium had to assist by providing the correct national versions from their own central archives.

To prevent this type of problem in future studies, the central archive at ACER was improved to host copies of all final national versions of the materials used in PISA 2006.

TRANSLATION AND VERIFICATION OUTCOMES – NATIONAL VERSION QUALITY

Analyses at the country level

One way to analyse the quality of a national version consists of analysing the item-by-country interaction coefficient. As the cognitive data have been scaled with the Rasch model for each country and for many languages (see Chapter 9), the relative difficulty of an item for a language within a country can be denoted δ_{ijk}, with *i* denoting the item, *j* denoting the language and *k* denoting the country. Further, each item can also be characterised by its international relative difficulty, denoted $\delta_{i\bullet\bullet}$, computed on a student random sample of equal size from all OECD country samples.

As both the national and international item calibrations were centred at zero, the mean of the δ_{ijk}, for any language *j* within a country *k* is equal to zero. In other words:

5.1

$$\sum_{i=1}^{I}\delta_{ijk} = 0 \quad \text{for all } j \text{ and } k$$

The item-by-country interaction is defined as the difference between any δ_{ijk} and its corresponding international item difficulty $\delta_{i\bullet\bullet}$. Therefore, the sum (and consequently the arithmetic mean) of the item-by-country interaction for a particular language within a country is equal to zero. Indeed,

5.2

$$\sum_{i=1}^{I}\left(\delta_{ijk} - \delta_{i..}\right) = \sum_{i=1}^{I}\delta_{ijk} - \sum_{i=1}^{I}\delta_{i..} = 0$$

As summary indices of item-by-country interaction for each language in a country we use the mean absolute deviation;

5.3

$$MAD_{jk} = \frac{1}{I}\sum_{i=1}^{I}\left|\delta_{ijk} - \delta_{i\bullet\bullet}\right|$$

and the root mean squared error

5.4

$$RMSE_{jk} = \sqrt{\frac{1}{I}\sum_{i-1}^{I}\left(\delta_{ijk} - \delta_{i\bullet\bullet}\right)^2}$$

and a chi-square statistic equal to;

5.5

$$X^2 = \sum_{i=1}^{I}\frac{\left(\delta_{ijk} - \delta_{i\bullet\bullet}\right)^2}{\mathrm{var}\left(\delta_{ijk}\right)}$$

As the sets item-by-country interactions by language and country, have a mean of zero, the mean of the absolute values is equal to the mean deviation and the root mean squared error is equal to the standard deviation of the item-by-country interactions.

A few science items were deleted at the national level (*i.e. S447Q02*, *S447Q03*, *S465Q04*, *S495Q04*, *S519Q01*, *S131Q04T*, *S268Q02T*, *S437Q03*, *S466Q01*, *S519Q03*, and *S524Q07*). To ensure the comparability of the analyses reported below, these items were removed from the science item parameter database and the national and international parameter estimates of the 92 remaining science items were re-centred on zero for each language and country.

Table 5.3 [Part 1/2]
Mean deviation and root mean squared error of the item by country interactions for each version

		Language	Absolute Value Mean or Mean deviation	RMSE or STD	X^2
OECD	Australia	English	0.24	0.29	223.99
	Austria	German	0.25	0.32	148.33
	Belgium	Dutch	0.28	0.34	173.36
	Belgium	French	0.25	0.31	110.95
	Belgium	German	0.25	0.32	59.60
	Canada	English	0.24	0.30	248.14
	Canada	French	0.20	0.28	118.04
	Czech Republic	Czech	0.25	0.32	156.73
	Denmark	Danish	0.22	0.30	133.23
	Finland	Finnish	0.34	0.43	235.97
	Finland	Swedish	0.38	0.51	80.94
	France	French	0.34	0.42	274.92
	Germany	German	0.25	0.31	142.98
	Greece	Greek	0.30	0.38	213.42
	Hungary	Hungarian	0.32	0.41	233.67
	Iceland	Icelandic	0.30	0.37	167.13
	Ireland	English	0.29	0.39	206.61
	Italy	German	0.30	0.38	110.40
	Italy	Italian	0.24	0.29	253.40
	Japan	Japanese	0.40	0.51	405.92
	Luxembourg	French	0.25	0.32	67.43
	Luxembourg	German	0.26	0.32	128.64
	Mexico	Spanish	0.31	0.40	580.70
	Netherlands	Dutch	0.30	0.39	217.46
	New Zealand	English	0.27	0.33	163.26
	Norway	Norwegian	0.23	0.30	130.45
	Poland	Polish	0.25	0.32	162.04
	Portugal	Portuguese	0.29	0.36	194.93
	Korea	Korean	0.42	0.55	433.22
	Slovak Republic	Hungarian	0.38	0.48	65.40
	Slovak Republic	Slovak	0.27	0.33	157.42
	Spain	Basque	0.37	0.47	136.18
	Spain	Catalan	0.28	0.35	103.32
	Spain	Galician	0.27	0.34	59.07
	Spain	Spanish	0.23	0.28	202.13
	Sweden	Swedish	0.23	0.29	121.16
	Switzerland	French	0.22	0.29	104.20
	Switzerland	German	0.25	0.31	188.76
	Switzerland	Italian	0.30	0.38	65.26
	Turkey	Turkish	0.32	0.41	247.18
	United Kingdom	English	0.29	0.36	291.11
	United Kingdom	Welsh	0.38	0.48	87.40
	United States	English	0.26	0.31	154.83

Table 5.3 [Part 2/2]
Mean deviation and root mean squared error of the item by country interactions for each version

		Language	Absolute Value Mean or Mean deviation	RMSE or STD	X^2
Partners	Argentina	Spanish	0.27	0.35	157.96
	Azerbaijan	Azeri	0.72	0.96	1115.60
	Azerbaijan	Russian	0.58	0.79	236.88
	Brazil	Portuguese	0.32	0.43	365.22
	Bulgaria	Bulgarian	0.29	0.38	209.40
	Chile	Spanish	0.26	0.32	166.02
	Colombia	Spanish	0.32	0.40	213.79
	Croatia	Croatian	0.30	0.40	225.32
	Estonia	Estonian	0.37	0.48	285.39
	Estonia	Russian	0.35	0.44	139.65
	Hong Kong-China	Chinese	0.45	0.56	418.56
	Indonesia	Indonesian	0.48	0.64	829.06
	Israel	Arab	0.41	0.51	156.82
	Israel	Hebrew	0.36	0.45	265.56
	Jordan	Arab	0.41	0.54	495.76
	Kyrgyzstan	Kyrgyz	0.62	0.79	526.08
	Kyrgyzstan	Russian	0.38	0.49	188.29
	Kyrgyzstan	Uzbek	0.64	0.79	238.67
	Latvia	Latvian	0.32	0.42	220.49
	Latvia	Russian	0.34	0.42	148.36
	Liechtenstein	German	0.36	0.46	76.65
	Lithuania	Lithuanian	0.37	0.47	323.31
	Lithuania	Russian	0.42	0.52	79.04
	Macao-China	Chinese	0.39	0.51	345.97
	Macao-China	English	0.46	0.57	155.65
	Montenegro	Montenegrin	0.37	0.45	291.95
	Qatar	Arab	0.47	0.57	425.06
	Qatar	English	0.45	0.58	241.25
	Romania	Hungarian	0.49	0.67	98.69
	Romania	Romanian	0.33	0.42	263.34
	Russian Federation	Russian	0.34	0.42	281.31
	Serbia	Hungarian	0.46	0.59	69.03
	Serbia	Serbian	0.30	0.40	233.18
	Slovenia	Slovenian	0.31	0.39	250.28
	Chinese Taipei	Chinese	0.51	0.66	839.30
	Thailand	Thai	0.38	0.48	385.94
	Tunisia	Tunisian	0.39	0.50	360.92
	Uruguay	Spanish	0.25	0.33	159.98

Country interactions for each language version are shown in Table 5.3. The six national versions with the highest mean deviation are:

- The Azeri version from Azerbaijan;
- The Uzbek version from Kyrgyzstan;
- The Kyrgyz version from Kyrgyzstan;
- The Russian version from Azerbaijan;
- The Hungarian version from Romania;
- The Chinese version from Chinese Taipei.

In a large number of countries with more than one language, the mean deviations of the different national versions are very similar. For instance, in Belgium, the mean deviations are respectively equal to 0.28, 0.25 and 0.25 for the Flemish version, the French version and the German version. In Estonia, they are respectively equal to 0.35 and 0.37 for the Estonian version and the Russian version. In Qatar, the English version and the Arabic version have a mean deviation of 0.45 and 0.47 respectively.

However, the mean deviations are quite different in a few countries. In Azerbaijan and in Kyrgyzstan, the mean deviation of the Russian version is substantially lower than the other national versions. The Hungarian versions used in Serbia, Romania and in the Slovak Republic present a larger mean deviation than the other national versions.

These results seem to indicate two sources of variability: the country and the language. The following tables present the correlations between the national version item parameter estimates for a particular language as well as the correlations between these item parameter estimates and the international item parameter estimates. If a language effect was suspected, then the within language correlations would be higher than the correlations with the international item parameter estimates.

Table 5.4
Correlation between national item parameter estimates for Arabic versions

	Israel	Jordan	Qatar	International Item Parameter
Israel				0.82
Jordan	0.84			0.82
Qatar	0.84	0.82		0.81
Tunisia	0.83	0.77	0.84	0.83

Table 5.5
Correlation between national item parameter estimates for Chinese versions

	Hong Kong-China	Macao-China	International Item Parameter
Hong Kong-China			0.82
Macao-China	0.94		0.85
Chinese Taipei	0.81	0.88	0.75

Table 5.6.
Correlation between national item parameter estimates for Dutch versions

	Belgium	International Item Parameter
Belgium		0.93
Netherlands	0.94	0.92

Table 5.7
Correlation between national item parameter estimates for English versions

	Australia	Canada	Great Britain	Ireland	Macao-China	New Zealand	Qatar	International Item Parameter
Australia								0.95
Canada	0.96							0.95
Great Britain	0.94	0.93						0.93
Ireland	0.92	0.93	0.96					0.92
Macao-China	0.77	0.80	0.80	0.79				0.80
New Zealand	0.98	0.95	0.94	0.91	0.77			0.94
Qatar	0.76	0.74	0.77	0.73	0.71	0.74		0.78
United States	0.97	0.96	0.94	0.91	0.78	0.95	0.81	0.94

Table 5.8
Correlation between national item parameter estimates for French versions

	Belgium	Canada	Switzerland	France	International Item Parameter
Belgium					0.94
Canada	0.95				0.95
Switzerland	0.97	0.96			0.95
France	0.94	0.90	0.94		0.89
Luxembourg	0.94	0.93	0.95	0.90	0.95

Table 5.9
Correlation between national item parameter estimates for German versions

	Austria	Belgium	Switzerland	Germany	Italy	Liechtenstein	International Item Parameter
Austria							0.95
Belgium	0.96						0.95
Switzerland	0.97	0.96					0.95
Germany	0.98	0.96	0.97				0.95
Italy	0.97	0.95	0.96	0.95			0.93
Liechtenstein	0.93	0.92	0.97	0.94	0.92		0.92
Luxembourg	0.96	0.96	0.97	0.97	0.96	0.93	0.95

Table 5.10
Correlation between national item parameter estimates for Hungarian versions

	Hungary	Romania	Serbia	International Item Parameter
Hungary				0.92
Romania	0.83			0.79
Serbia	0.89	0.81		0.85
Slovak Republic	0.93	0.80	0.87	0.89

Table 5.11
Correlation between national item parameter estimates for Italian versions

	Italy	International Item Parameter
Italy		0.95
Switzerland	0.95	0.92

Table 5.12
Correlation between national item parameter estimates for Portuguese versions

	Brazil	International Item Parameter
Brazil		0.88
Portugal	0.87	0.94

Table 5.13
Correlation between national item parameter estimates for Russian versions

	Azerbaijan	Estonia	Kyrgyzstan	Lithuania	Latvia	International Item Parameter
Azerbaijan						0.65
Estonia	0.76					0.89
Kyrgyzstan	0.81	0.88				0.86
Lithuania	0.79	0.89	0.84			0.85
Latvia	0.76	0.95	0.89	0.89		0.89
Russia	0.80	0.96	0.92	0.90	0.95	0.89

Table 5.14
Correlation between national item parameter estimates for Spanish versions

	Argentina	Chile	Colombia	Spain	Mexico	International Item Parameter
Argentina						0.93
Chile	0.94					0.94
Colombia	0.92	0.91				0.90
Spain	0.93	0.92	0.90			0.96
Mexico	0.92	0.92	0.93	0.90		0.91
Uruguay	0.94	0.93	0.91	0.93	0.93	0.93

Table 5.15
Correlation between national item parameter estimates for Swedish versions

	Finland	International Item Parameter
Finland		0.90
Sweden	0.94	0.95

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

For the various Arabic-, Dutch-, German- and Spanish-language versions, the within-language correlations do not differ substantially from the correlations between the national and the international item parameter estimates.

The correlations within the Chinese-language versions are substantially higher than their respective correlations with the international item parameter estimates. This might reflect a language effect or a cultural effect, included a curriculum effect.

The correlations within English-language versions show an interesting pattern. First of all, the correlations between parameter estimates for the English-language versions from the two countries where English is a minority language (*i.e.* Qatar and Macao-China) are lower than the respective correlations for the countries where English is the majority language. Further, the English-speaking countries seem to form two groups: Great Britain and Ireland in the first group and the others in the second group. Within a group, the correlations between the national versions are higher than their correlations with the international items parameter estimates while between group, the correlations appears to be equal or lower than the correlations with the international item parameter estimates.

The correlation pattern of the French-language versions outlines an increase of the correlation for France. While the item parameter estimates for France correlate at 0.89 with the international item parameter estimates, they correlate at 0.94 with the item parameter estimates of the French-language version of Belgium and Switzerland.

The Hungarian-language versions from Romania, Serbia and the Slovak Republic better correlate with the national version of Hungary than with the international item parameter estimates. The same phenomenon is also observed with the Russian-language versions. For any country that tested some part of their population in the Russian language, the item parameter estimates correlate better with the item parameter of Russia than with the international item parameter estimates.

Table 5.16
Correlation between national item parameter estimates within countries

		Language 1	Language 2	Correlation
OECD	Belgium	Dutch	French	0.89
		Dutch	German	0.89
		French	German	0.90
	Canada	English	French	0.92
	Switzerland	French	German	0.91
		French	Italian	0.93
		German	Italian	0.92
	Spain	Basque	Catalan	0.87
		Basque	Galician	0.89
		Basque	Spanish	0.91
		Catalan	Galician	0.93
		Catalan	Spanish	0.94
		Galician	Spanish	0.95
	Finland	Finish	Swedish	0.86
	Slovak Republic	Slovak	Hungarian	0.87
	United Kingdom	English	Welsh	0.89
Partners	Azerbaijan	Russian	Azeri	0.77
	Estonia	Estonian	Russian	0.85
	Israel	Hebrew	Arabic	0.81
	Kyrgyzstan	Uzbek	Kyrgyz	0.90
		Kyrgyz	Russian	0.84
		Uzbek	Russian	0.82
	Lithuania	Russian	Lithuanian	0.78
	Latvia	Russian	Latvian	0.89
	Macao-China	English	Chinese	0.78
	Qatar	English	Arabic	0.91
	Romania	Romanian	Hungarian	0.83
	Serbia	Serbian	Hungarian	0.80

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Among all these correlation matrices, it appears that the matrix for the English version is the most instructive. It seems that the cultural effects or the curriculum effect are more important than the language effects. To confirm this hypothesis, correlations have been computed between national versions within countries. If the hypothesis is correct, then the correlation between the national versions within a country should be higher than the correlation between national versions within languages.

Based on Table 5.16, a few observations can be made:

- Where a country has borrowed a version from another country or if countries have cooperated to produce a common version, the national item parameter estimates better correlates within the language than within the country. For instance, the Belgian-Flemish version shows a higher correlation with the Dutch version than with the Belgian-French version. This is also the case for the Swedish version in Finland;
- As the correlation between the national item parameter estimates of the two versions in Canada (English and French) is lower than most of the correlations for the English version and the French version, one cannot dismiss some effect of the language;
- The correlation between the Arabic-language Qatari version the three national versions in Kyrgyzstan seem to reflect a curriculum effect. While the English-language version and the Arabic-language version in Qatar correlate respectively at 0.78 and 0.80 with the international item parameter estimates, they correlate 0.91 with each other. Also, while the Kyrgyz-language version and the Uzbek-language version correlate respectively 0.73 and 0.69 with the international item parameter estimates, they correlate 0.90 with each other;
- On the other hand, for Macao-China, the correlation between different language versions is not higher than the correlation with the international item parameter estimates. This could reflect some translation or equivalence issues.

To further disentangle the effects, variance decomposition models of the absolute value of the item-by-country interaction have been performed.

Table 5.17 shows the results of a nested analysis of variance of the absolute value of the item by country interaction of the 92 science items, which includes those countries with multiple language versions and the multiple versions for each country are treated as nested within the country.

Table 5.17
Variance estimate

	Variance estimates	Variance estimates without Azerbaijan and Kyrgyzstan
Country	0.010	0.003
Version (Country)	0.003	0.002
Residual	0.090	0.069

The country variance estimate is substantially higher than the version-within-country variance estimate. However, as already mentioned, Azerbaijan and Kyrgyzstan national versions had high mean deviations and low correlation with the international item parameter estimates. Without these two countries, the country variance estimates and the version-within-country variance estimates are quite similar. In each case, the most important variance component is the residual. To better understand the meaning of this residual, the unit and the item effects were included in the decomposition of the item by country interactions.

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Table 5.18
Variance estimates

Effect	Variance estimate
Test unit	0.00095
Item within unit	0.00279
Country	0.00317
Country by test unit	0.00132
Country by item within unit	0.00632
Version within country	0.00226
Version within country by test unit	0.00002
Version within country by item within unit	0.05783

Table 5.18 presents the variance decomposition with four main effects, (i) country, (ii) language version nested in country, (iii) test unit and (iv) item embedded nested unit. Science units with a single item and countries with only one national version were therefore removed from the database. It therefore remains 17 countries, 38 countries representing 23 languages, 87 items embedded in 31 units.

The first two variance estimates are a test effect. They both reflect that some units, on average, have more item-by-country interactions than others and more particularly that some items have on average larger item-by-country interactions than others. The next section of this chapter is devoted to analyses at the item and the unit levels.

The second set of variance estimates provided in Table 5.18 are cultural or curriculum effects. The country effect,in Table 5.3 confirms that some countries have on average, larger item-by-country interactions than others. The interaction between the country and the unit reflects that some units are relatively easier or more difficult for the different national versions within a country. Finally, the interaction between the country and the item, which is the largest effect after the residual effect, confirms that some items appear to be relatively easier or more difficult for the different versions within a country. As it is quite unlikely that a translation problem occurs for the same unit or for the same item in each national version within a country, and further has the same effect, these two interactions can therefore be considered as cultural effect or curriculum effect.

Finally, the last three effects show equivalence problems, translation problems or a cultural and/or curriculum, linguistic effect. Indeed, in countries like Belgium, there are no national curricula, as education is a responsibility of the linguistic communities.

About 75% of the variability of the item-by-country interaction is at the lowest level, *i.e.* the interaction between the item and the national version.

Analyses at the item level

On average across countries, a unit has an item-by-country interaction of 0.34. It ranges from 0.25 for unit *S447* to 0.44 for unit *S493*. None of the unit characteristics (*i.e.* application area, original language of the item) are related to the unit item-by-country interaction average.

The average item-by-country interaction at the item level ranges from 0.19 (*S498Q04*) to 0.53 (*S458Q01*). The item format and the item focus do not affect the item-by-country interaction. average but the assessed competency is significantly associated with the item-by-country interaction. Items designed for assessing *using scientific evidence* on average present a mean item-by-country interaction of 0,33, items for *identifying scientific issues* a mean of 0.33 and items for *explaining phenomena scientifically* a mean of 0.36.

Summary of items lost at the national level, due to translation, printing or layout errors

In all cases when large DIF or other serious flaws were identified in specific items, the NPMs were asked to review their translation of the item and to provide the consortium with possible explanations.

As often happens in this kind of exercise, no obvious translation error was found in a majority of cases. However, some residual errors could be identified, that had been overlooked by both the NPMs and the verifier. Out of the 179 mathematics, reading and science items, 28 items were omitted in a total of 38 occurrences for the computation of national scores for the following reasons:

- Mistranslations or confusing translations: 20 items;
- Poor printing: 13 items;
- Layout issues: one item;
- Omission of key words: three items;
- Problematic item since PISA 2000: one item.

Reader's Guide

Country codes – the following country codes are used in this report:

OECD countries

AUS	Australia
AUT	Austria
BEL	Belgium
BEF	Belgium (French Community)
BEN	Belgium (Flemish Community)
CAN	Canada
CAE	Canada (English Community)
CAF	Canada (French Community)
CZE	Czech Republic
DNK	Denmark
FIN	Finland
FRA	France
DEU	Germany
GRC	Greece
HUN	Hungary
ISL	Iceland
IRL	Ireland
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
LXF	Luxembourg (French Community)
LXG	Luxembourg (German Community)
MEX	Mexico
NLD	Netherlands
NZL	New Zealand
NOR	Norway
POL	Poland
PRT	Portugal
SVK	Slovak Republic
ESP	Spain
ESB	Spain (Basque Community)
ESC	Spain (Catalonian Community)
ESS	Spain (Castillian Community)
SWE	Sweden
CHE	Switzerland
CHF	Switzerland (French Community)
CHG	Switzerland (German Community)
CHI	Switzerland (Italian Community)
TUR	Turkey
GBR	United Kingdom
IRL	Ireland
SCO	Scotland
USA	United States

Partner countries and economies

ARG	Argentina
AZE	Azerbaijan
BGR	Bulgaria
BRA	Brazil
CHL	Chile
COL	Colombia
EST	Estonia
HKG	Hong Kong-China
HRV	Croatia
IDN	Indonesia
JOR	Jordan
KGZ	Kyrgyztan
LIE	Liechtenstein
LTU	Lithuania
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
MAC	Macao-China
MNE	Montenegro
QAT	Qatar
ROU	Romania
RUS	Russian Federation
SRB	Serbia
SVN	Slovenia
TAP	Chinese Taipei
THA	Thailand
TUN	Tunisia
URY	Uruguay

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

References

Adams, R.J., Wilson, M. & **Wang, W.C.** (1997), The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* No. 21, pp. 1-23.

Adams, R.J., Wilson, M. R. & **Wu, M.L.** (1997), Multilevel item response models: An approach to errors in variables regression, *Journal of Educational and Behavioural Statistics,* No. 22 (1), pp. 46-75.

Adams, R.J. & **Wu, M.L.** (2002), *PISA 2000 Technical Report,* OECD, Paris.

Bollen, K.A. & **Long, S.J.** (1993) (eds.), *Testing Structural Equation Models*, Newbury Park: London.

Beaton, A.E. (1987), Implementing the new design: The NAEP 1983-84 technical report (Rep. No. 15-TR-20), Princeton, NJ: Educational Testing Service.

Buchmann, C. (2000), Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school? *Soc. Forces,* No. 78, pp. 1349-79.

Buchmann, C. (2002), Measuring Family Background in International Studies of Education: Conceptual Issues and Methodological Challenges, in Porter, A.C. and Gamoran, A. (eds.). *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 150-97), Washington, DC: National Academy Press.

Creemers, B.P.M. (1994), *The Effective Classroom*, London: Cassell.

Cochran, W.G. (1977), *Sampling techniques,* third edition, New York, NY: John Wiley and Sons.

Ganzeboom, H.B.G., de Graaf, P.M. & **Treiman, D.J.** (1992), A standard international socio-economic index of occupational status, *Social Science Research*, No. 21, pp. 1-56.

Ganzeboom H.B. & **Treiman, D.J.** (1996), Internationally comparable measures of occupational status for the 1988 international standard classification of occupations, *Social Science Research*, No. 25, pp. 201-239.

Grisay, A. (2003), Translation procedures in OECD/PISA 2000 international assessment, *Language Testing,* No. 20 (2), pp. 225-240.

Hambleton, R.K., Swaminathan, H. & **Rogers, H.J.** (1991), *Fundamentals of item response theory*, Newbury Park, London, New Delhi: SAGE Publications.

Hambleton, R.K., Merenda, P.F. & **Spielberger, C.D.** (2005), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment,* IEA Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.

Harkness, J.A., Van de Vijver, F.J.R. & **Mohler, P.Ph** (2003), *Cross-Cultural Survey Methods,* Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey.

Harvey-Beavis, A. (2002), Student and School Questionnaire Development, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report,* (pp. 33-38), OECD, Paris.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Jöreskog, K.G. & **Sörbom, Dag** (1993), *LISREL 8 User's Reference Guide*, Chicago: SSI.

Judkins, D.R. (1990), Fay's Method of Variance Estimation, *Journal of Official Statistics*, No. 6 (3), pp. 223-239.

Kaplan, D. (2000), *Structural equation modeling: Foundation and extensions,* Thousand Oaks: SAGE Publications.

Keyfitz, N. (1951), Sampling with probabilities proportionate to science: Adjustment for changes in probabilities, *Journal of the American Statistical Association,* No. 46, American Statistical Association, Alexandria, pp. 105-109.

Kish, L. (1992), Weighting for Unequal, *Pi. Journal of Official Statistics*, No. 8 (2), pp. 183-200.

LISREL (1993), K.G. Jöreskog & D. Sörbom, [computer software], Lincolnwood, IL: Scientific Software International, Inc.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxberry: Pacific Grove.

Macaskill, G., Adams, R.J. & **Wu, M.L.** (1998), Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scale, in M. Martin and D.L. Kelly, Editors, *Third International Mathematics and Science Study, technical report Volume 3: Implementation and analysis*, Boston College, Chestnut Hill, MA.

Masters, G.N. & **Wright, B.D.** (1997), The Partial Credit Model, in W.J. van der Linden, & R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory* (pp. 101-122), New York/Berlin/Heidelberg: Springer.

Mislevy, R.J. (1991), Randomization-based inference about latent variables from complex samples, *Psychometrika,* No. 56, pp. 177-196.

Mislevy, R.J., Beaton, A., Kaplan, B.A. & **Sheehan, K.** (1992), Estimating population characteristics from sparse matrix samples of item responses, *Journal of Educational Measurement,* No. 29 (2), pp. 133-161.

Mislevy, R.J. & **Sheehan, K.M.** (1987), Marginal estimation procedures, in Beaton, A.E., Editor, 1987. *The NAEP 1983-84 technical report,* National Assessment of Educational Progress, Educational Testing Service, Princeton, pp. 293-360.

Mislevy, R.J. & **Sheehan, K.M.** (1989), Information matrices in latent-variable models, *Journal of Educational Statistics,* No. 14, pp. 335-350.

Mislevy, R.J. & **Sheehan, K.M.** (1989), The role of collateral information about examinees in item parameter estimation, *Psychometrika,* No. 54, pp. 661-679.

Monseur, C. & **Berezner, A.** (2007), The Computation of Equating Errors in International Surveys in Education, *Journal of Applied Measurement,* No. 8 (3), 2007, pp. 323-335.

Monseur, C. (2005), An exploratory alternative approach for student non response weight adjustment, *Studies in Educational Evaluation,* No. 31 (2-3), pp. 129-144.

Muthen, B. & **L. Muthen** (1998), [computer software], *Mplus* Los Angeles, CA: Muthen & Muthen.

Muthen, B., du Toit, S.H.C. & **Spisic, D.** (1997), *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes,* unpublished manuscript.

OECD (1999), *Classifying Educational Programmes. Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2003), *Literacy Skills for the World of Tomorrow: Further results from PISA 2000,* OECD, Paris.

OECD (2004), *Learning for Tomorrow's World – First Results from PISA 2003,* OECD, Paris.

OECD (2005), *Technical Report for the OECD Programme for International Student Assessment 2003,* OECD, Paris.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006,* OECD, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World,* OECD, Paris.

PISA Consortium (2006), *PISA 2006 Main Study Data Management Manual, https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_data_management_manual.pdf*

Rasch, G. (1960), Probabilistic models for some intelligence and attainment tests, Copenhagen: Nielsen & Lydiche.

Routitski A. & **Berezner, A.** (2006), Issues influencing the validity of cross-national comparisons of student performance. Data Entry Quality and Parameter Estimation. Paper presented at the Annual Meeting of the American Educational Research Association (AERA) in San Francisco, 7-11 April, *https://mypisa.acer.edu.au/images/mypisadoc/aera06routitsky_berezner.pdf*

Rust, K. (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics,* No. 1, pp. 381-397.

Rust, K.F. & **Rao, J.N.K.** (1996), Variance Estimation for Complex Surveys Using Replication Techniques, *Survey Methods in Medical Research,* No. 5, pp. 283-310.

Shao, J. (1996), Resampling Methods in Sample Surveys (with Discussion), *Statistics,* No. 27, pp. 203-254.

Särndal, C.-E., Swensson, B. & **Wretman, J.** (1992), *Model Assisted Survey Sampling,* New York: Springer-Verlag.

SAS® CALIS (1992), W. Hartmann [computer software], Cary, NC: SAS Institute Inc.

Scheerens, J. (1990), School effectiveness and the development of process indicators of school functioning, *School effectiveness and school improvement,* No. 1, pp. 61-80.

Scheerens, J. & **Bosker, R.J.** (1997), *The Foundations of School Effectiveness,* Oxford: Pergamon.

Schulz, W. (2002), Constructing and Validating the Questionnaire composites, in R.J. Adams and M.L. Wu (eds.), *PISA 2000 Technical Report,* OECD, Paris.

Schulz, W. (2004), Mapping Student Scores to Item Responses, in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study, Technical Report* (pp. 127-132), Amsterdam: IEA.

Schulz, W. (2006a), *Testing Parameter Invariance for Questionnaire Indices using Confirmatory Factor Analysis and Item Response Theory,* Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Schulz, W. (2006b), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003,* Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Thorndike, R.L. (1973), *Reading comprehension in fifteen countries,* New York, Wiley: and Stockholm: Almqvist & Wiksell.

Travers, K.J. & **Westbury, I.** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula,* Oxford: Pergamon Press.

Travers, K.J., Garden R.A. & **Rosier, M.** (1989), Introduction to the Study, in Robitaille, D. A. and Garden, R. A. (eds), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula,* Oxford: Pergamon Press.

Verhelst, N. (2002), Coder and Marker Reliabilaity Studies, in R.J. Adams & M.L. Wu (eds.), *PISA 2000 Technical Report*. OECD, Paris.

Walberg, H.J. (1984), Improving the productivity of American schools, *Educational Leadership,* No. 41, pp. 19-27.

Walberg, H. (1986), Synthesis of research on teaching, in M. Wittrock (ed.), *Handbook of research on teaching* (pp. 214-229), New York: Macmillan.

Walker, M. (2006), *The choice of Likert or dichotomous items to measure attitudes across culturally distinct countries in international comparative educational research*. Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Walker, M. (2007), Ameliorating Culturally-Based Extreme Response Tendencies To Attitude items, *Journal of Applied Measurement,* No. 8, pp. 267-278.

Warm, T.A. (1989), Weighted Likelihood Estimation of Ability in Item Response Theory, *Psychometrika*, No. 54 (3), pp. 427-450.

Westat (2007), *WesVar® 5.1* Computer software and manual, Rockville, MD: Author (also see *http://www.westat.com/wesvar/*).

Wilson, M. (1994), Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity, in M. Wilson (ed.), *Objective measurement II: Theory into practice* (pp. 271-292), Norwood, NJ: Ablex.

Wolter, K.M. (2007), *Introduction to Variance Estimation*. Second edition, Springer: New York.

Wu, M.L., Adams, R.J. & **Wilson, M.R.** (1997), *ConQuest®: Multi-Aspect Test Software* [computer program manual], Camberwell, Vic.: Australian Council for Educational Research.

List of abbreviations – the following abbreviations are used in this report:

ACER	Australian Council for Educational Research
AGFI	Adjusted Goodness-of-Fit Index
BRR	Balanced Repeated Replication
CBAS	Computer Based Assessment of Science
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ENR	Enrolment of 15-year-olds
ESCS	PISA Index of Economic, Social and Cultural Status
ETS	Educational Testing Service
IAEP	International Assessment of Educational Progress
I	Sampling Interval
ICR	Inter-Country Coder Reliability Study
ICT	Information Communication Technology
IEA	International Association for the Evaluation of Educational Achievement
INES	OECD Indicators of Education Systems
IRT	Item Response Theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index
MENR	Enrolment for moderately small school
MOS	Measure of size
NCQM	National Centre Quality Monitor
NDP	National Desired Population
NEP	National Enrolled Population
NFI	Normed Fit Index
NIER	National Institute for Educational Research, Japan
NNFI	Non-Normed Fit Index
NPM	National Project Manager
OECD	Organisation for Economic Cooperation and Development
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
PGB	PISA Governing Board
PQM	PISA Quality Monitor
PSU	Primary Sampling Units
QAS	Questionnaire Adaptations Spreadsheet
RMSEA	Root Mean Square Error of Approximation
RN	Random Number
SC	School Co-ordinator
SE	Standard Error
SD	Standard Deviation
SEM	Structural Equation Modelling
SMEG	Subject Matter Expert Group
SPT	Study Programme Table
TA	Test Administrator
TAG	Technical Advisory Group
TCS	Target Cluster Size
TIMSS	Third International Mathematics and Science Study
TIMSS-R	Third International Mathematics and Science Study – Repeat
VENR	Enrolment for very small schools
WLE	Weighted Likelihood Estimates

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

Table of contents

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

LIST OF BOXES

LIST OF FIGURES

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

LIST OF TABLES

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009

PISA 2006 TECHNICAL REPORT – ISBN 978-92-64-04808-9 – © OECD 2009