JOURNAL OF APPLIED MEASUREMENT, *8*(3), 249-266


# Translation Equivalence across PISA Countries

Aletta Grisay
*University of Liège, Belgium*

John H.A.L. de Jong
*Language Testing Services, Netherlands*

Eveline Gebhardt
*Australian Council for Educational Research, Melbourne*

Alla Berezner
*Australian Council for Educational Research, Melbourne*

Beatrice Halleux-Monseur
*HallStat SPRL, Belgium*

Due to the continuous increase in the number of countries participating in international comparative assessments such as TIMSS and PISA, ensuring linguistic and cultural equivalence across the various national versions of the assessment instruments has become an increasingly crucial challenge. For example, 58 countries participated in the PISA 2006 Main Study. Within each country, the assessment instruments had to be adapted into each language of instruction used in the sampled schools. All national versions in languages used for 5 per cent or more of the target population (that is, a total of 77 versions in 42 different languages) were verified for equivalence against the English and French source versions developed by the PISA consortium. Information gathered both through the verification process and through empirical analyses of the data are used in order to adjudicate whether the level of linguistic equivalence reached an acceptable standard in each participating country.

The paper briefly describes the procedures typically used in PISA to ensure high levels of translation/adaptation accuracy, and then focuses on the development of the set of indicators that are used as criteria in the equivalence adjudication exercise. Empirical data from the PISA 2005 Field Trial are used to illustrate both the analyses and the major conclusions reached.

Requests for reprints should be sent to Aleta Grisay, 22, Avenue des Canadiens, F-94410 Saint-Maurice, France, e-mail: agrisay@attglobal.net.

## Introduction

All OECD countries except Turkey and the Slovak Republic participated in the first PISA assessment (PISA 2000), and all 30 of them participated in the PISA 2003 and PISA 2006 studies (that is, Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States of America). While these countries differ widely in terms of their geographic, linguistic and cultural characteristics, they are probably more homogeneous, particularly on economic and social grounds, than the groups of countries participating in other international studies, such as TIMSS.

Since the commencement of PISA, however, a progressively increasing number of non-OECD partner economies have joined the study (five in PISA 2000, eight in the PISA 2000 replication of 2001, eleven in PISA 2003, and 27 in PISA 2006).[1] The consequence has been that the number of languages into which the assessment instruments have to be translated has increased from 25 to 42 languages and the number of national versions to be independently checked for equivalence with the source versions has increased from 41 to 77. This has added considerably to the diversity, and to the challenge of ensuring equivalence and fairness of the instruments across all participating countries.

Strict procedures were implemented in PISA in order to ensure the development of high quality adaptations of the instruments in all languages of instruction used in the participating countries, as well as independent verification of their equivalence with the source versions produced by the International PISA Centre (IPC).

The typical PISA procedures include, in particular, the development of two parallel source versions (in English and French), with a recommendation that each country develops two separate versions in their language of instruction (one from each source language), then reconciles them into a final national version (Grisay, 2003a).

Both the English and French source versions provided by the PISA IPC include frequent translation notes aimed at helping with possible translation or adaptation problems. For example, with each Reading item the objective of the item is explicitly stated to further the likelihood that translated items set similar requirements (McQueen and Mendelovits, 2003). In addition, a comprehensive document describing the recommended translation procedure and containing detailed translation/adaptation guidelines is 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 are then submitted for central verification against the source versions to a group of international verifiers appointed and specially trained by the PISA IPC. This verification team is composed of professional translators proficient in both English and French, and with native command of the target language. After entering the corrections proposed by their verifier (or sometimes discussing and rejecting a few of them, or finding alternative solutions), the participating countries are asked to return hard copies of their future test booklets, so that the verifier can perform a final check on the accurateness of edits, the correct assembly of the material, the layout and the graphics.

All participating countries are asked to establish a National Expert Committee, which has the responsibility of reviewing the appropriateness of source material for the country's 15-year-old

[1] Partner economies in PISA 2000 were Brazil, Hong Kong, Indonesia, Latvia, the Russian Federation and Thailand. The additional partner economies that participated in PISA2000 follow-up were Albania, Argentina, Bulgaria, Chile, Israel, the Former Yugoslavian Republic of Macedonia, Peru and Romania. The 11 partner economies involved in PISA 2003 were Brazil, Hong Kong, Indonesia, Latvia, Liechtenstein, Macao, the Russian Federation, the Former Yugoslavian Republic of Serbia, Thailand, Tunisia and Uruguay. In PISA 2006, the 27 partner economies included Argentina, Azerbaijan, Brazil, Bulgaria, Chile, Colombia, Croatia, Estonia, Hong Kong, Indonesia, Israel, Jordan, Kyrgyz Republic, Latvia, Liechtenstein, Lithuania, Macao, the Former Yugoslavian Republics of Serbia and Montenegro, Qatar, Romania, the Russian Federation, Slovenia, Taiwan, Thailand, Tunisia and Uruguay.

students (in particular by checking the items for potential inconsistencies with the national curriculum). The committee assists national translators with terminology and other content-specific problems, and reviews and endorses the final national version.

Both at the Field Trial and Main Study phase of each cycle, all countries receive from the IPC a detailed report based on the item analysis of their national data, including, in particular, a "dodgy items list" pointing at items that had item/country interaction or at other types of flaws in their data set. National centres are asked to review these items; when plausible explanations are found for a specific item, it is either corrected (at the Field Trial phase) or, in some cases, deleted from the analysis (at the Main Study phase).

Finally, in each cycle, at the end of the Main Study data analysis phase, a Data Adjudication expert panel composed of PISA analysts, of Technical Advisory Group members and of domain experts is in charge of assessing whether each of the participating countries met the PISA technical standards and whether therefore their results can be recommended for inclusion in the PISA reports. A document describing these standards has been endorsed by the PISA Governing Board and was circulated to all national centres (OECD, 2007).

In PISA 2000 and PISA 2003, the information about translation/adaptation that was provided to the experts involved in the Data Adjudication exercise was mainly drawn (i) from the operational reports received from the national Centres (e.g., did they use professional translators? Did they train them using the PISA guidelines? Did they implement a double-translation and reconciliation procedure? Did they submit their materials for verification? Did they have their materials reviewed by a national expert committee?) and, (ii) from qualitative information provided in the verifiers' reports (main types and frequency of errors encountered).

In addition, a study was conducted of all PISA 2000 national versions that used one of the three languages shared by a significant number of participating countries, i.e., English, French and German (Grisay, 2003b). The item difficulty parameter estimates were used to identify those cross-country differences in the item difficulties that appeared to be common to all or most countries in one of the language groups (by contrast to the other two groups), suggesting that possible translation flaws might have had some impact on the item behavior in one of these languages.

However, the data provided in PISA 2000 and 2003 did not allow for formal reporting on translation equivalence, according to two of the Test Adaptation Guidelines defined by the International Test Commission (Hambleton, 1994; Hambleton and Merenda, 2005):

1. Guideline D5: "Test developers…should implement systematic judgmental evidence—both linguistic and psychological—to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions," and
2. Guideline D9: "Test developers… should provide statistical evidence of the equivalence of questions for all intended populations."

In the PISA 2006 Field Trial, a much more systematic review of equivalence issues was conducted. As suggested by Zumbo (2003) two categories of analyses were employed: item-level analyses, and scale-level analyses.

At the item level, Le (2006) conducted a comprehensive exploration of Differential Item Functioning (DIF) by country, by language, and by gender, in relation to the Science item characteristics as defined in the assessment framework—item format, type of context and scientific competence assessed.

On the other hand, at the whole scale level, this article explores the development of a small set of potential indicators, which might help with the internal management of the translation and verification activities, and with formal reporting on translation equivalence.

In terms of internal management, it should be possible to use the indicators to help in checking, for example: (i) whether the two source versions could be considered as equivalent at the highest possible level of construct invariance (Van de

Vijver, 1998), and equally free of flawed or biased items; (ii) whether certain national versions appear to be of poor quality and to necessitate direct action (for example, examining whether certain translators or certain international verifiers need to be replaced); and, (iii) whether the quality of the translated materials tends to improve (or to deteriorate) from one cycle to another in certain countries.

In terms of formal reporting, the indicators should provide information to be used in the adjudication of countries' data, to report on whether there were any countries with national versions too severely flawed for their materials to be considered as *equivalent* when compared to the source versions and the majority of other national versions.

This article presents and discusses the analyses that were conducted in a first tentative step towards the development of the desired set of indicators.

*Identifying geographic and linguistic patterns of differences using cluster analysis*

An effective method for identifying systematic patterns in the functioning of various versions of a test instrument used in an international study is conducting a cluster analysis of the differences in item difficulty indices observed across the national versions.

This method was used, for example, by Blum, Goldstein and Guérin-Pace (2001) in a reanalysis of the IALS data, in order to show that the hierarchies of items, ranked according to the proportion of correct answers in each participating country, differed significantly from one country to another. In particular, all of the English-speaking countries participating in IALS were grouped in one of the clusters, while the French-speaking countries and the German-speaking countries formed separate clusters. The authors concluded that "*…the item success rate [was] associated with geographic and linguistic factors, which [contradicted] the hypothesis of comparability which underpins this survey, based on the assumption that performance is independent of the language of questioning."*

Similar analyses, with similar results, were conducted in PISA 2000, PISA 2003 and PISA 2006—but rather than the raw percentages of correct answers, the authors used the item difficulty parameters (deltas) derived from independent Rasch analyses conducted for each of the national data subsets. In Figure 1 the dendrogram obtained from a cluster analysis of the item deltas observed in the PISA 2006 Field Trial Science test is presented.[2]

The patterns of similarities and differences evidenced in Figure 1, like those in the IALS data, reveal linguistic, geographic and cultural factors.

With very few exceptions, the national versions sharing the same language appeared in the same cluster. This was the case, for example, for most of the versions in German, in English, in French, in Dutch, in Spanish, in Russian, in Chinese and in Arabic. The main dendrogram structure, by itself, seemed to be related to a linguistic typology, where the same large cluster associated all Germanic languages (i.e., German, English, Dutch and the group of Scandinavian languages), another cluster regrouped the Romance languages (i.e., French, Italian, Portuguese, Spanish and the various Spanish dialects), and still another grouped the Slavic and Baltic languages (Russian, Polish, Serb, Croatian, Montenegrin, Slovenian, Slovak, Lithuanian and Latvian).

However, a clear impact of geographic proximity can also be perceived. For example, the three versions in Finno-Ugrian languages were included in three quite different "geographic" clusters: the Finnish version among the German and Scandinavian versions, the Estonian version in a small Baltic group and the Hungarian version in a loose Eastern European group that mainly included Slavic languages, but also one Romance language, Rumanian. Similarly, the La-

[2] This analysis was based on 70 Field Trial data sets (one data set for each of the languages used within each of the countries) and a pool of 201 Science items retained in the Field Trial analyses after dropping some 46 items that had severe flaws in both the source versions and a large number of participating countries – thus suggesting that the weaknesses were with the items themselves rather than with their translation.

tino-American Spanish and Portuguese versions appeared in the same subgroup, independent from the European versions in Portuguese, Spanish, Catalan and Galician. The Hebrew and Arabic versions used in Israel were in the same subgroup, despite the linguistic difference.

Conversely, the English version used in Qatar was fairly distant from the Arabic version used in the same country and from the two other Arabic versions used in the study (Jordan and Tunisia). And the Russian version used in Kyrghystan was part of the Russian subgroup (together with the Russian versions used in Russia, in Latvia and Estonia), fairly distant from the version in Kyrghyz used in the same country. This may indicate that in some cases, the use of minority languages in certain countries could be associated with different curricula or/and very different population subgroups.

Finally some impact of the countries' cultural and socio-economic characteristics can also be detected. For example, the most developed Asian countries (Japan, Korea, Hong Kong, Chinese Taipei and Macao) are included in the same cluster, which tends to be somewhat more similar to the large group of Western and Indo-European clusters than to the most "distant" group of participating countries, which includes two less industrialized Pacific countries (Thailand and Indonesia), and a number of Arabic-speaking countries (Qatar (Arabic), Jordan, Tunisia as well as all countries using Turkish/Altaic languages (Turkey, Kyrghystan (Uzbek and Kyrghyz) and Azerbaijan (Azeri)). This latter group is mainly characterized by a much lower GDP per capita than in the average OECD country, and by relatively large proportions of students who perform at the lowest levels of the PISA proficiency scales.

Only minimal changes were observed in the cluster structure when the analysis was replicated using only the subset of items that were eventually retained by the test developers and by the Science Expert group for use in the PISA 2006 Main study. The patterns resulting from this analysis appeared therefore to convey relatively robust information on the fact that the Science materials used in the PISA 2006 Field Trial did not function exactly in the same manner in all participating countries, and that the differences in languages used probably played a role in the item/country interactions which were causing the patterns observed.

*Assessing the magnitude of possible linguistic bias*

However, the dendrogram presented in Figure 1 does not contain information on the *magnitude* of the effects related to linguistic differences. Many of the patterns observed could well be due to minor differences affecting, for example, just two or three items in a specific group of countries—with negligible impact on the overall comparability of the Science scale at the international level.

A factor analysis was therefore used to estimate the amount of variance in item difficulties that was common across the subsets of data corresponding to the various national versions, as compared to the variance explained by secondary factors or to specificities that were unique to single versions or groups of versions. The covariance matrix used in the factor analysis had 70 variables (the national versions) and 201 observations (the Field Trial items retained by the test developers after deleting the group of items that had severe content problems).

The first factor extracted by the analysis explained 79.2 per cent of the total variance; two thirds of the versions had loadings of more than 0.90 on that factor, and for no version the loading was less than 0.74, indicating substantial comparability of the construct across all countries.

Two additional factors had eigenvalues larger than one. The second (unrotated) factor explained a further 2.7 per cent of the variance and had modest positive loadings for a number of versions used in Islamic countries (Azerbaijan, Kyrgyzstan, Turkey, Indonesia, Jordan, Tunisia, Qatar) but zero or small negative loadings for all other versions. The third factor (1.6 per cent of the variance) was uninterpretable.

It must be noted that the same type of analysis, conducted by Baye (2004) on data from the PISA 2000 Reading Main study, also extracted

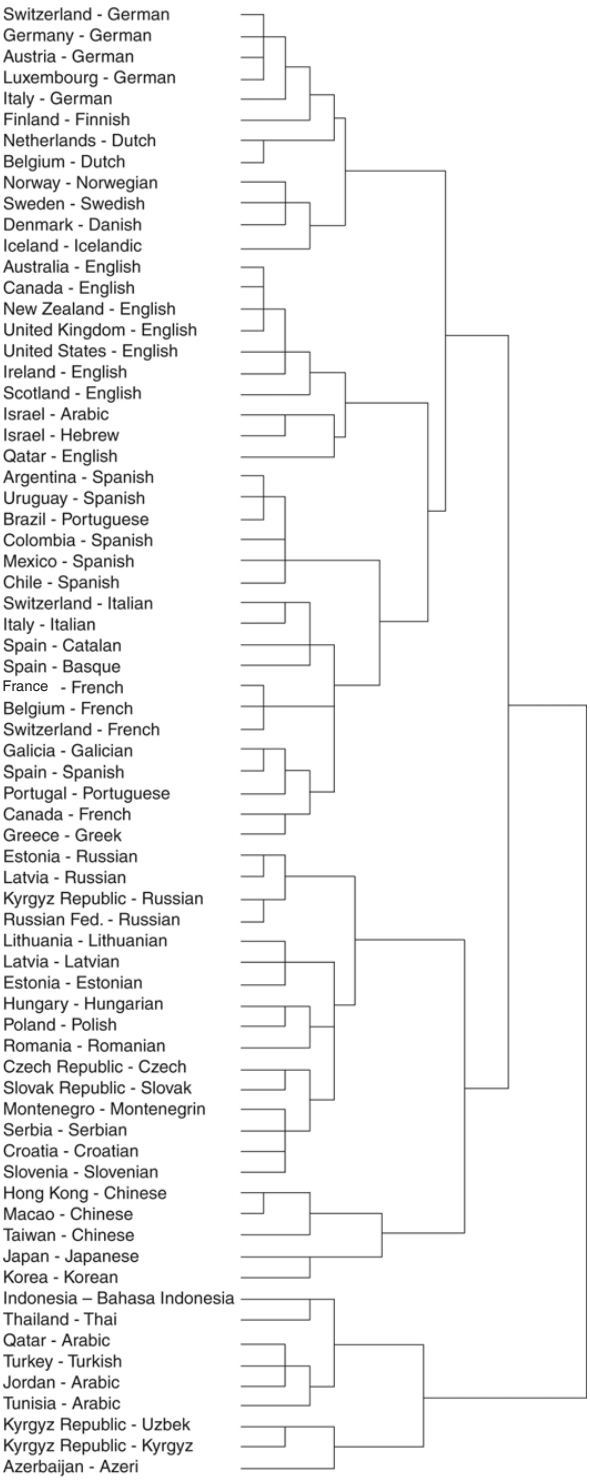


*Figure 1*. Cluster Analysis of Item Difficulties across National Versions

a very strong first factor (82 per cent of the total variance), and two minor and uninterpretable other factors. We repeated the analysis using data from the PISA 2003 Mathematics Main Study, and the first factor accounted for 91 per cent of the total variance. In both cases the dendrogram produced by a cluster analysis of item difficulties showed linguistic and geographic dependencies very similar to those in Figure 1. This would seem to indicate that linguistic and cultural specificities are consistent and play some role in all three domains, although the variance components involved appear as relatively minor, compared to the variance attributable to the common latent dimension measured by the test across all participating countries.

A first indicator of equivalence between the national versions of the Science test used in the PISA 2006 Field Trial was therefore constructed using the communalities provided by a one-factor PCA, and has been presented in Figure 2.

In Figure 2, the light grey section of each bar indicates the communality for a given national version obtained from the factor analysis, i.e., the *variance in item difficulties that is shared with all other national versions*. The remaining (unique) variance is split up in two parts: *random error* on one hand (white section), and on the other hand a variance component potentially caused by *various sources of bias* (black section).

The white section corresponds to the *amount of random error*, mainly due to the size of the samples of students used in the Field Trial and to other *modelled* sources of random error. In particular, the versions used for minority groups of students in a number of multilingual countries had been usually administered to smaller samples than those in the dominant language, and had, as a consequence, significantly larger amounts of random error. The amount of random error was estimated using a simulation that included the following steps:

1. Since each student took about 50 items in the PISA 2006 booklet rotation design, 50 *true* item difficulties were generated from a normal distribution.

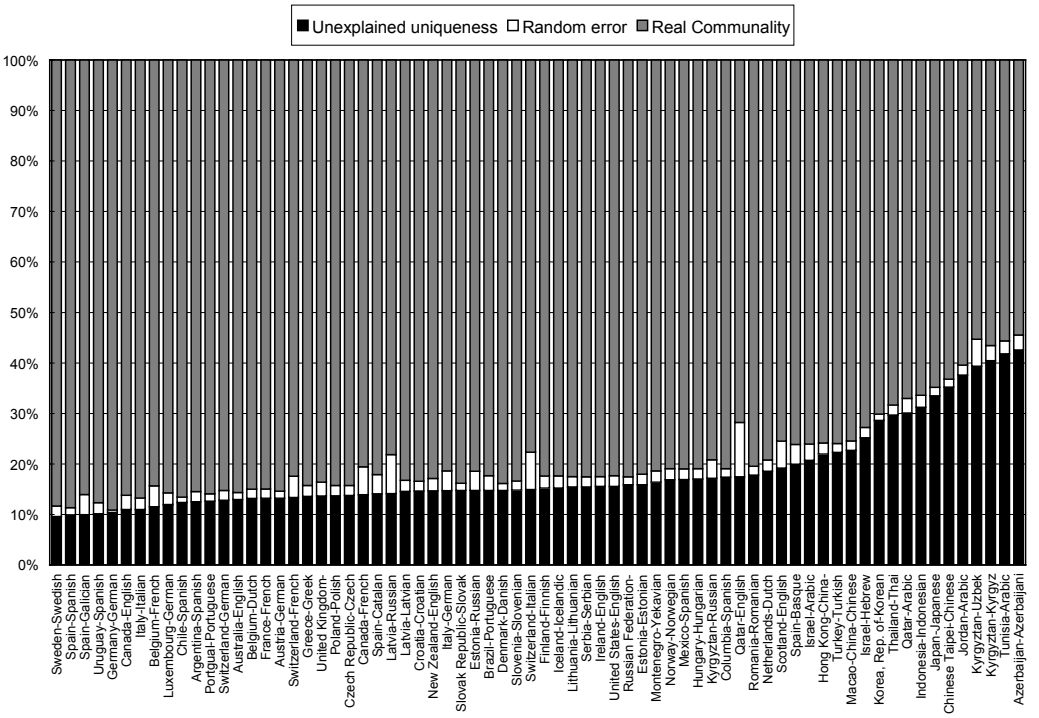


*Figure 2*. Communalities and uniqueness in science parameter items by country

2. Student responses were generated for each of the country-language combinations, using the *true* item parameters, the mean performance of these countries (from the field trial), the variance in performance within countries, and the sample size (average number of students per item).
3. These item responses were analyzed and item parameters were estimated. So these national item parameters were slightly different from the true item parameters because of (a) random error, (b) targeting, (c) sample size, and (d) variance in performance within the country.
4. The national item parameters derived from the simulation were used in a one-factor PCA analysis. In this simulation the unique variance was completely *caused* by random error, targeting, sample size and probably to a lesser extent the variance of student scores within a country.
5. Steps 2 to 4 were repeated 50 times, and the value of the error components for each country was averaged over the 50 simulations, in order to obtain accurate estimations. The error component was then subtracted from the *unique variance* left in the real data after extracting the common factor, thus eliminating the part that is caused by random error, by between- and within-countries variance, and by sample size. The average proportion of error variance was 2.4 per cent, but the estimated amount varied from near zero in Germany (where each booklet had been administered to about 1400 students), up to 11 per cent for the English version used in Qatar, (that had only 60-70 students per booklet).

The black section is the unexplained component that remains after subtracting the error component (as described in step 5) from the actual unique variance in item difficulty associated with each national version. It can be considered as a tentative indicator of the part of variance in a country's data that is neither common to the international scale, nor attributable to purely random factors, and is therefore likely to be due to *bias affecting the equivalence with other versions*.

From Figure 2, it can be concluded that the proportion of bias variance was small and relatively uniform across more than two thirds of the national versions. In particular, there was no evidence in these data that the English and French national versions directly derived from the source versions had significantly less bias than those developed through translation and adaptation from the two source versions into other Western or European languages.

However, a significant group of national versions, mainly used in Middle East and Asian countries, appearing at the bottom of the graphic, showed quite high values of the "uniqueness" indicator (from 20 per cent of the variance for the Arabic version used in Israel up to 43 per cent for the (independent) Arabic version used in Tunisia and for the Azeri version). Three of the versions in this group were used in OECD Member countries (Turkey, Japan and Korea), and the countries involved included both some of those where the students' proficiency in Science was the highest in the PISA 2006 Field Trial (Chinese Taipei, Japan, Korea, Hong Kong) and some of those where it was the lowest (Azerbaijan, Kyrgyzstan, Qatar, Indonesia, Thailand, Tunisia, Turkey, Jordan).

*Understanding some of the potential sources of bias*

In order to better understand the reasons why the PISA Science instrument seemed to behave in a less "equivalent" way in this group of participating countries than in others, a number of other indicators describing countries' characteristics of potential interest were developed, and used in a multiple regression analysis in order to "predict" the magnitude of the indicator of unique variance. These included:

- *A proxy for linguistic and cultural differences conveyed by the language of instruction.* A dichotomized variable opposing the Indo-European languages *vs* the Non-Indo-European languages was retained, after a few attempts at using a more detailed classification. The correlation between the indicator of uniqueness and information on the language of instruction did not improve significantly when

using specific categories such as Germanic, Romance, Slavic, Altaic, Finno-Ugrian languages, instead of the main contrast between Indo-European *vs* non-Indo-European languages.

- *A proxy for possible economic differences.* The country's GDP per capita, expressed in US dollars at purchasing power parity, was used to represent possible differences due to the relative level of economic development in the participating countries. When a country had more than one national version, the same GDP value was imputed to all versions used in the country.
- *A proxy for possible differences in the quality of translation*, derived from the verifiers' reports. In PISA 2006, the verifiers appointed by the IPC had been requested to systematically report all errors found in the national versions that they had identified as a threat for the equivalence against the source versions. These errors were described (in English) in Excel spreadsheets that were then submitted for adjudication to the International PISA Centre. Based on these reports, the IPC earmarked the errors considered as "key issues" that requested correction before the use of the national version in the Field Trial data collection could be authorised. The proxy variable used was the number of "key corrections" requested per 100 000 characters of text in each national version. The indicator was far from perfect, since, by definition, most of these errors were corrected before the Field Trial, and also because the global number of "key issues" submitted to the IPC was dependent on the accuracy and on the personal judgement of more than 40 different verifiers. The hypothesis behind this indicator was, however, that those national versions where large numbers of serious translation errors had been identified and corrected were more likely than other versions to contain residual translation errors that had escaped the verifiers' vigilance.
- *A proxy for possible differences in Science curriculum coverage,* derived from the national review of the Field Trial items received from each National Centre. The National Project Directors had been requested to assess the relevance of each of the Science items for their national curriculum on a scale from 1 (not included in the national curriculum used for 15 years old students) to 5 (perfectly appropriate item, given the curriculum taught to a vast majority of 15 years old students in the country). The indicator used was the average value of this variable across the 247 Science items used in the Field Trial in each country. When a country had more than one national version, the same curriculum coverage value was imputed to all versions used in the country.
- *A proxy for possible differences in the quality of the national coding for open-ended questions.* A homogeneity analysis was used to estimate the coding consistency among the national staff. The analysis used a sample of student answers for each of the open-ended items in the PISA 2006 Field Trial test, which were coded independently by four different markers. High values of the homogeneity index correspond to high between-coders agreement.[3]
- *A proxy for possible targeting effects,* over and above those resulting in random errors already dealt with in the simulation described above. Ceiling or floor effects may have resulted in poorer discrimination coefficients for groups of items at the extremes of the distribution of item difficulties in countries with particularly high or low mean scores. The indicator retained to represent this potential source of bias was the average index of discrimination of the Science items for the group of students who were administered each of the versions of the Field Trial test.

In Table 1 the relationships within this group of variables and between them and the indica-

[3] Note that this index does not capture possible coding *bias* (i.e., cases when the codes from the various markers are consistently more lenient or more harsh than required by the international coding manuals). At the Main Study phase, an additional quality control exercise is conducted using independent markers appointed by the International PISA Centre, from which separate indicators of consistency of coding from country to country are derived.

Table 1

*Correlations between the indicator of unexplained variance and proxy variables for potential sources of bias*

| Variable | Uniqueness indicator | Mean science score | Field trial sample size | Non–Indo–European language | Country's GDP | No. of key translation corrections per 100000 char. | Curriculum coverage | Within-country coding reliability index | Average index of Item Discrim. |
|---|---|---|---|---|---|---|---|---|---|
| Uniqueness indicator | 1.00 | | | | | | | | |
| Mean science score | –0.23 | 1.00 | | | | | | | |
| Field trial sample size | –0.01 | 0.00 | 1.00 | | | | | | |
| Non Indo–European lang. | 0.75 | –0.29 | 0.08 | 1.00 | | | | | |
| Country's GDP | –0.36 | 0.61 | –0.21 | –0.32 | 1.00 | | | | |
| No of key transl. corrections | 0.20 | –0.09 | 0.08 | 0.38 | –0.27 | 1.00 | | | |
| Curriculum coverage | –0.16 | 0.09 | 0.04 | –0.18 | 0.05 | 0.05 | 1.00 | | |
| Coders reliability index | 0.05 | –0.15 | 0.11 | 0.06 | –0.13 | –0.15 | –0.25 | 1.00 | |
| Average Item Discrim. | –0.51 | 0.40 | 0.02 | –0.46 | 0.59 | –0.44 | 0.03 | –0.01 | 1.00 |

tor of uniqueness in item difficulties have been presented.

The index of *Coders' reliability* appeared to have no significant relation with the indicator of *Uniqueness* nor with any of the other variables, and therefore this indicator was not used in further analyses.

The correlation between the indicator of *Uniqueness* and the proxies for *Curriculum coverage* and for quality of translation (*Key Corrections*) had the expected direction: the amount of bias variance was proportionally lesser in countries where the curriculum taught to the 15 years old students covered most of the topics assessed in the PISA Science test, and for versions where the verifiers identified fewer translation errors. However, these correlations were very modest, suggesting that the two variables measured in a less than perfect way these potential sources of bias.

When *Key Corrections* and *Curriculum Coverage* were included in a stepwise regression where *Uniqueness* was the dependent variable (see Table 2), it appeared that their contribution was almost entirely confounded with that of *GDP* and of the dichotomized variable *Non-Indo-European* language, indicating that, to some extent, the amount of non-equivalence was associated with a combination of factors: many countries with non Indo-European languages had lower GDP than other countries; their translations may have been somewhat poorer and their curriculum did not cover all of the topics assessed. This combination of characteristics was associated with higher levels of bias.

In fact, as can be seen from the regression coefficients in Table 2 only *Non Indo-European language* and *Average Item Discrimination* contributed significant amounts of unique variance in the last step of the analysis. By alternating the order of the variables entered in the various steps, the variance of the indicator of bias in the national versions could be approximately decomposed as follows: unique variance explained by *Non Indo-European language,* 24%; unique variance explained by *Average Item Discrimination,* 10%; variance contributed by *Curriculum Coverage* and

Table 2

*Stepwise regression of the indicator of Uniqueness in item difficulties on the various proxy variables*

| Predictors included | Increase in $R^2$ | Cumulative $R^2$ |
|---|---|---|
| Step 1. Curriculum coverage | 0.07 | 0.07 |
| Step 2. As above, plus Key corrections | 0.08 | 0.15 |
| Step 3. As above, plus GDP | 0.12 | 0.27 |
| Step 4. As above, plus Non Indo–European language | 0.34 | 0.61 |
| Step 5. As above, plus Average Item Discrimination | 0.10 | 0.71 |

Regression Coefficients at Step 5

| | Estimate | *S.E.* | *T*–value | *Pr* > \|*t*\| |
|---|---|---|---|---|
| Intercept | 0.670 | 0.076 | 8.86 | <0.0001 |
| Curriculum coverage | –0.012 | 0.008 | –1.40 | 0.166 |
| Key corrections | –0.000 | 0.000 | –1.17 | 0.245 |
| GDP | 0.000 | 0.000 | 0.38 | 0.708 |
| Non Indo–European language | 0.104 | 0.015 | 7.10 | <0.0001 |
| Average Item Discrimination | –0.959 | 0.206 | –4.65 | <0.0001 |

Alternative Models

| Predictors included | Explained $R^2$ |
|---|---|
| A. Only Non Indo–European language | 0.56 |
| B. Only Average Item Discrimination | 0.43 |
| C. Both the above included | 0.69 |

*Key Corrections*, 2%; variance explained jointly by *Non Indo-European language* and by *Average Item Discrimination* (also partly confounded with *GDP*, *Key corrections* and *Curriculum coverage*), 35%; and, non-explained variance: 29%

This pattern of results is of some concern. It indicates that the amount of uniqueness in the distribution of item difficulties across the various national versions can be explained to a large extent (about 70 per cent of the variance of this indicator) by the larger distance between the international instruments and the national contexts in certain countries—in terms of linguistic and cultural characteristics of the instruments and their less than optimal targeting for the levels of students' performance in those countries.

In particular, the versions used in a number of low-GDP Middle-East and Asian non-OECD countries seemed to suffer both from cultural distance *and* from low item discrimination due to very high numbers of low-achieving students, which may explain the large joint variance component observed in the analysis above.

Interestingly, item discrimination did not seem to play any role for high-achieving Asian countries, such as Japan, Korea, Chinese Taipei and Hong Kong, but linguistic distance probably did, as confirmed indirectly by their relatively high numbers of translation corrections. Similarly, some of the national versions in non Indo-European languages that were used in the group of Western countries (Finnish, Hungarian, Estonian, and Basque) tended to have slightly higher values of the uniqueness indicator than other European countries, also probably related to linguistic issues.

Would the equivalence be improved if specific groups of items were selected for the Main Study?

A few additional analyses were conducted in order to check whether changing the composition of the set of items used in the test would have a significant impact on the communalities, particularly in countries that had large unique variances.

*Impact of item format*

First, two separate factor analyses were used to compute the communalities that would have been obtained if the PISA 2006 Science test

had contained only closed questions (Multiple Choice or Complex Multiple Choice items) or only open-ended questions (Short Constructed Responses or Complex Constructed Responses) (see Figure 3). About two thirds of the 201 Science test items (126 items) were included in the MC/CMC analysis, and the remaining third (75 items) was used in the Open-ended analysis. Figure 4 allows a comparison of the communalities obtained when using all items (squares), those obtained when using only MC or CMC items (circles), and those obtained when using only open-ended items (triangles).

Interestingly, the overall communality appeared to be slightly better for a test composed of MC or CMC items only (81 per cent of the total variance explained, as compared to 79 per cent in the factor analysis using the complete set of items), while the set of open-ended items produced in general lower communalities (explaining 75 per cent of the overall variance). As can be seen in Figure 3, the difference between the communalities resulting from the two sets of items was quite significant for some of the national versions (the Czech version, the Basque version used in Spain, both the English and Arabic versions used in Qatar, the Arabic versions used in Jordan and Tunisia, and, most of all, the Azeri version, (where the communality for MC and CMC items was near twice as large as that observed for Open-ended items).

This result is somewhat counter-intuitive. One would have expected that the student's answers to multiple-choice items might have been more affected by cultural and educational differences in familiarity with that item format than their responses to open-ended items. However, the latter are more dependent than multiple-choice items upon the quality of manual coding – which suggests that part of the bias observed in some of the countries with high values of the uniqueness indicator might perhaps be due to coding problems that were not measured by the current index of inter-coder reliability.

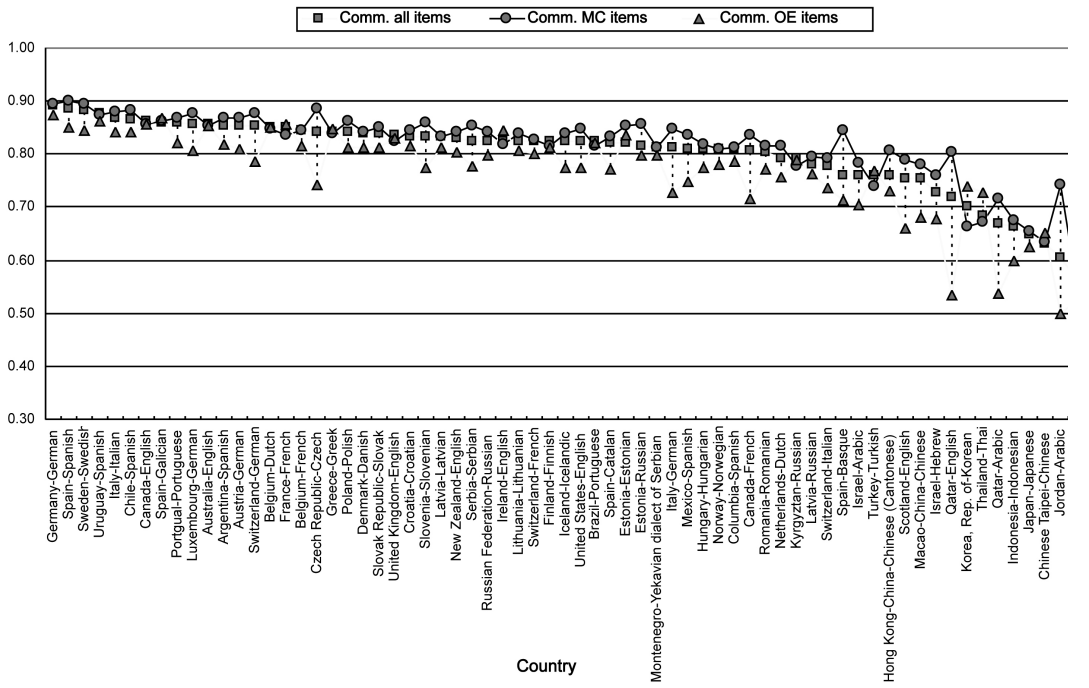


*Figure 3*. Communalities for potential tests composed of multiple-choice or open-ended items only

*Impact of item difficulty*

A second exploration, illustrated in Figure 4, consisted of running three separate factor analyses using (i) the 75 per cent easiest items from the Field Trial test; (ii) the 75 per cent most difficult items, and (iii) a set of 75 per cent items selected at random.

As can be observed in Figure 4, in all countries the "best" communalities were obtained when using a random mix of easy, hard and medium difficulty items. The common factor explained 79 per cent of the total variance in this analysis, as compared to 68 per cent in both the analysis including a majority of easy or medium difficulty items and the analysis including hard or medium difficulty items. This probably indicates that reducing or increasing the overall difficulty of the instrument would not improve the equivalence.

However, the graph in Figure 4 also suggests that, although the overall amount of common variance would be about the same for the "easy" and the "hard" instrument, their behavior would not be exactly the same across all national versions. In some countries (like Japan, Korea, Indonesia, Thailand, Scotland or Brazil), the "easy" instrument would produce lower communalities than its "hard" counterpart. The reverse seemed to be true for a number of other versions (Czech Republic, Russian Federation, Slovak Republic, Croatia, Hungary, Latvia (Russian), Argentina and Jordan).

To explore further potential interactions between item difficulties and mean achievement of the groups of students who were administered the various versions of the PISA 2006 Science Test, the average amount of positive or negative DIF that was observed in each version for the 25 per cent "easiest" and "hardest" items has been computed.

As a general rule, there was more instability with the easiest than with the hardest group of items. The easiest items appeared to function as if they were harder than expected in a majority

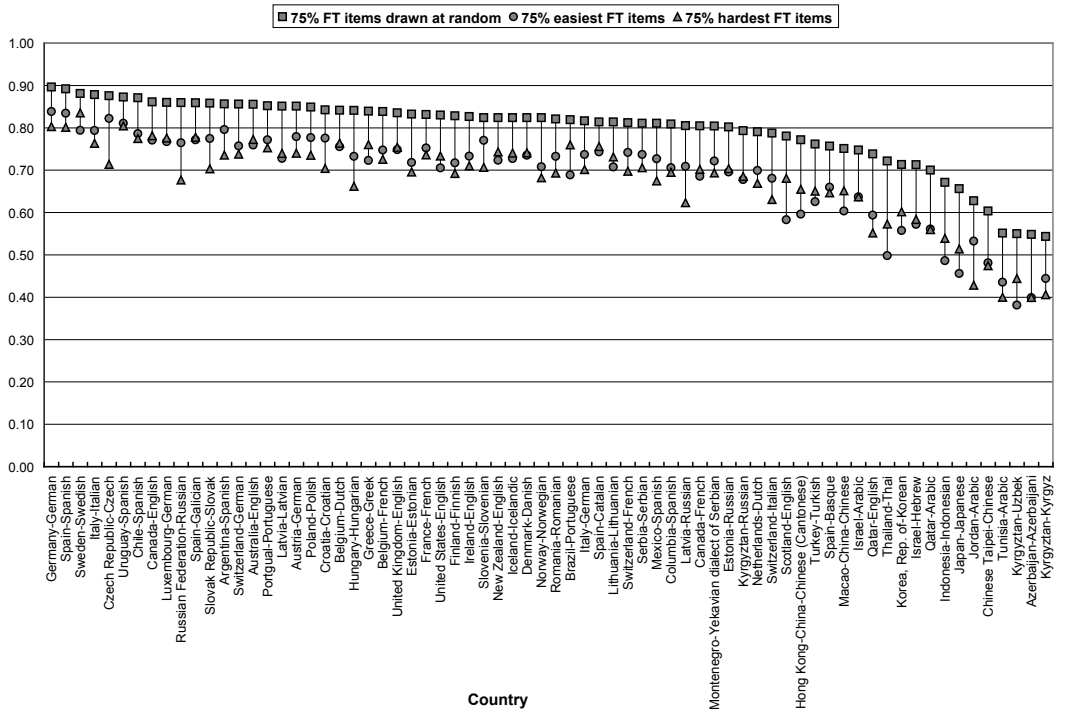


*Figure 4*. Communalities for potential tests composed of easier or harder items

of countries that obtained low average scores in the PISA 2006 Field Trial (except for Mexico), but also in Japan and Korea. By contrast, they were slightly easier than expected in Finland and Czech Republic.

A "mirror" trend was observed to some extent for the hardest items, which tended to appear as easier than expected in a number of low-achieving countries. However, the amount of both positive and negative DIF was globally lower for the group of hard items than for the easy ones.

*Comparing the equivalence indicators for PISA 2003 and TIMSS 2003 Mathematics assessments.*

Finally, an analysis was conducted using data from the Mathematics assessments conducted in 2003 both by PISA and TIMSS, in order to examine to what extent the pattern of results described above was specific to Science or common to both Science and Mathematics, and whether the lower communalities observed in non-Western countries in PISA also tended to appear in TIMSS or were mainly dependent on the particular group of countries participating in each study.

Since for TIMSS the information on item difficulties was available from the Web as average per cent of correct answers by item and by country, the same type of information was computed for the PISA items and used in the factor analyses instead of the delta parameters.

The results, presented in Figure 5, are interesting in many regards. First, the communalities observed in PISA 2003 for Mathematics appeared to be, on average, significantly higher than those reported in Figure 2 for Science in PISA 2006 (91.3 per cent of the total variance in Mathematics and 79.2 per cent in Science). The trend was similar in TIMSS 2003, with higher communalities for Mathematics than for Science in virtually all countries.[4] This suggests that the curriculum in Mathematics is probably more universal than in

[4] No direct comparison between the PISA and TIMSS communalities in Science is included in this paper, since the PISA data are only from a Field Trial.

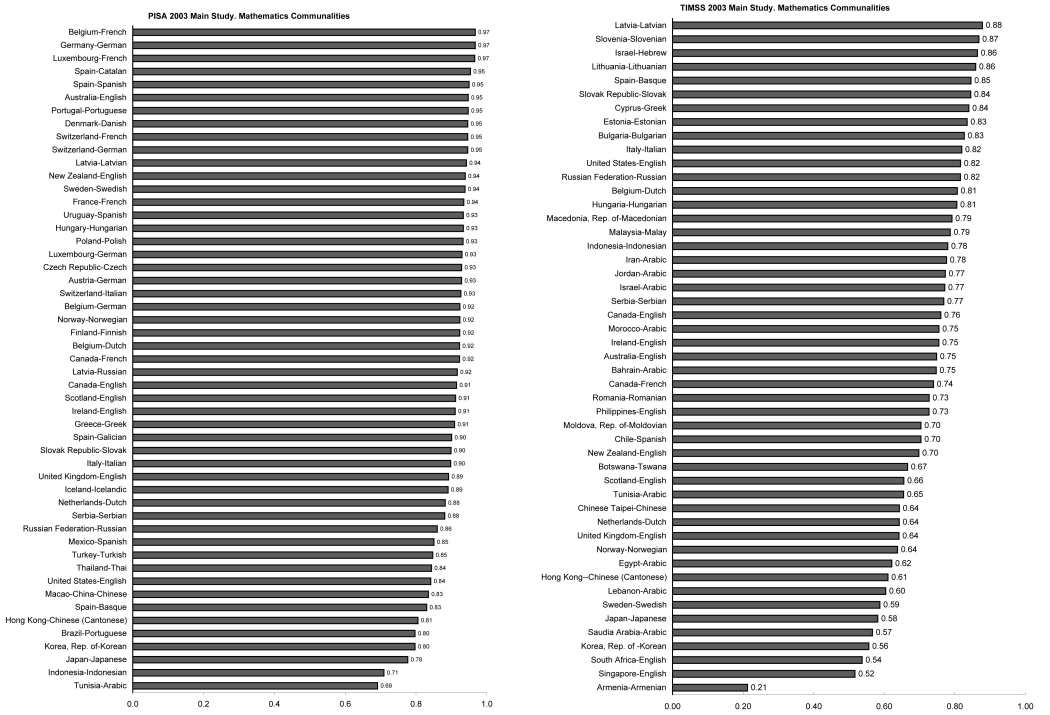


*Figure 5*. PISA and TIMSS 2003 Mathematics assessment. Communalities in item difficulties

Science, where the topics taught may differ more widely from country to country.

Second, and quite surprisingly, the Mathematics communalities were in general significantly lower in TIMSS than in PISA (on average, 71.9 per cent of the total variance in TIMSS and 91.3 per cent in PISA). Potential reasons for this large difference seemed to be (i) the higher cultural homogeneity of the group of countries participating in PISA, and, (ii) the more directly curriculum-dependent nature of the TIMSS instruments. As can be seen from Figure 6, the average difference in communalities between the two studies was reduced by half (but did not disappear: 77.5 per cent of common variance in TIMSS and 87.5 per cent in PISA) when the analysis was repeated using only the group of 22 countries that participated in both studies, and a selected set of TIMSS items containing exactly the same proportion of questions related to Number, Algebra, Geometry, Measurement and Data as in PISA.

Third, independently from this difference between the two studies, some similarity was observed in the patterns of distribution of communalities across European and non-European countries and across industrialized and developing countries. In TIMSS, like in PISA, the group of countries with the lowest indices of communality tended to be mainly Asian countries (Hong Kong, Japan, Chinese Taipei, Korea, Singapore) and part of the Arabic-speaking countries (Tunisia, Egypt, Lebanon, Saudi Arabia). For the group of 22 countries that participated in both studies, the correlation between the rank orders of communalities in PISA and TIMSS was 0.63, ($p < 0.001$).

## Discussion

Equivalence in an international test can be defined as an equal probability of getting any particular item correct for all students at a given level of proficiency, independent of the national version they were administered. Checking the order of item difficulties in each participating country against the mean percent correct values obtained at the international level has been common practice in a number of IEA studies, many years before the introduction of IRT models. Unfortunately, however, this information has not been routinely reported, and can therefore not be used for comparisons of the quality of the instruments used in various cross-national studies.

An exception was the IEA/Reading Literacy study: in an appendix of the international report, Elley (1992) noted that the mean correlation between the national and international rank order of item difficulties was approximately 0.92 for the Reading test administered to 9 years old students, and 0.91 for the test administered to 14 years old students. Commenting on these results, the author suggested that "while some minor features may still be found to exist which a few observers would perceive as lending a cultural bias," their impact was unlikely to be larger than the bias variance due to cultural factors in any national test. He concluded that the IEA/RLS tests results "are believed to be comparable across countries, as they would be within countries."

This optimism is in sharp contrast with the conclusions of a recent study conducted by Ercikan and Koh (2005), in which they examined the construct comparability of the English and French versions of TIMSS 1995 as used in England, Canada, France and the USA. CFA analyses were used to test a full measurement equivalence model (i.e., a model assuming invariance of number of factors, of item loadings, of item errors, and of correlations between factors across the different data sets) for each of the 8 Mathematics and Science test booklets used in TIMSS. The RMSEA values obtained indicated that there was a good fit of the full model to the data for 3 Maths booklets and one Science booklet in the Canadian (ENG) / Canadian (FRE) comparison and in the England / France comparison, but for none of the eight booklets in the US / France comparison, neither in Mathematics nor in Science.

In their conclusion Ercikan and Koh indicated that there were considerable differences assessed by the English and French versions of mathematics and sciences. They noted that "differences created by the adaptation process as well as linguistic differences that might affect

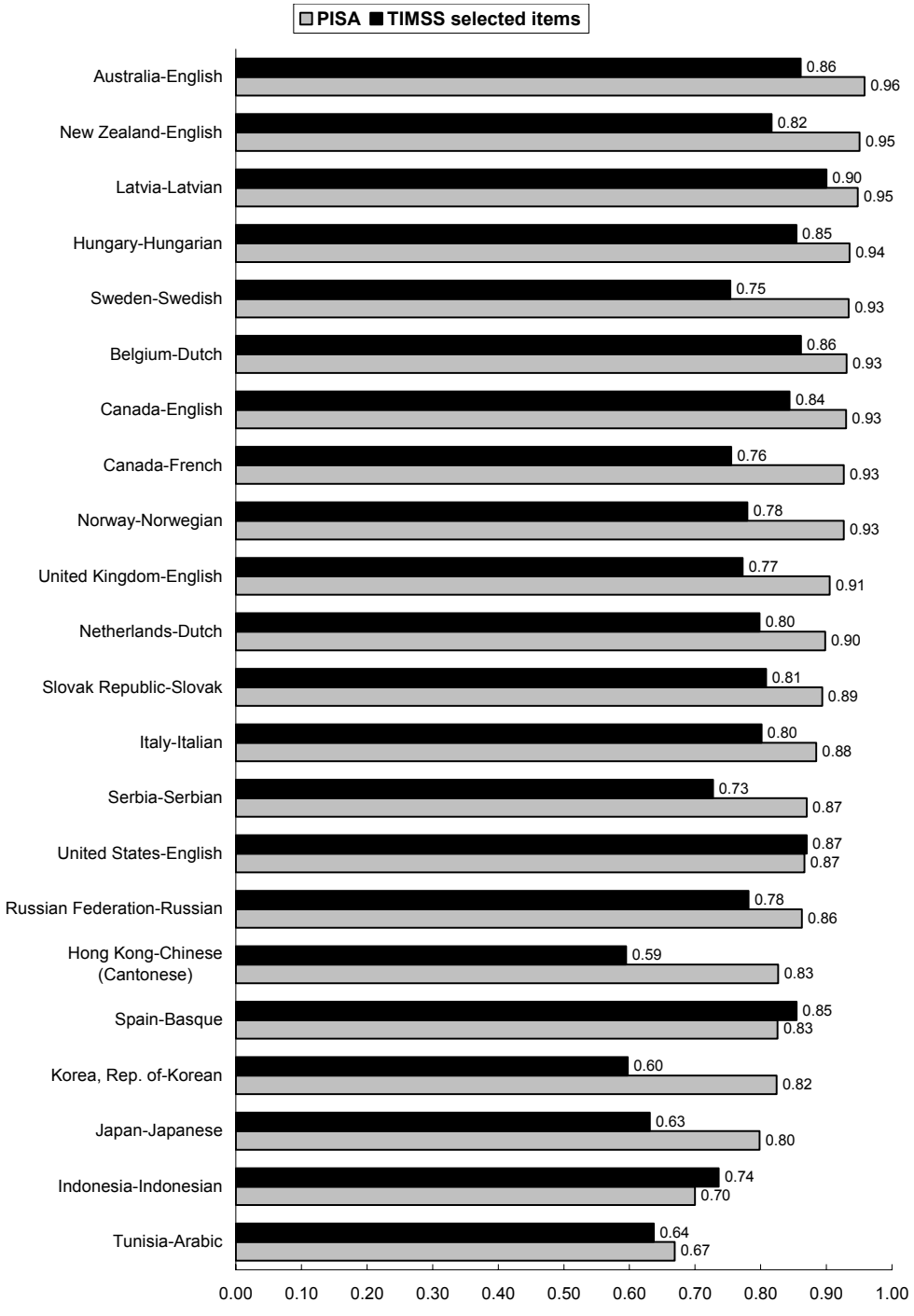


*Figure 6*. PISA 2003 and TIMSS 2003 Communalities of item difficulties for the 22 countries that participated in both studies

examinee performance can affect the equivalence of constructs assessed in different countries" and concluded that: "The results from this study point to differences in constructs assessed by TIMSS in different countries and the importance of empirical evidence to support construct comparability before TIMSS results can be meaningfully used for research."

While the analyses conducted in our own study were far less demanding than the CFA analyses used by Ercikan and Koh, they confirmed the need for more systematic reports on equivalence across national instruments in international studies.

In fact, the correlations mentioned by Elley (1992) were very similar to the corresponding coefficients computed for the PISA 2000 Reading test, indicating in both studies that in a vast majority of the participating countries the national versions of the instruments functioned in a very consistent way.

However, in both studies some of the lowest correlations were observed for the Asian participating countries (between 0.75 and 0.85 for Singapore, Hong Kong and Thailand in the IEA/RLS study, and for Hong Kong, Korea, Japan, Thailand and Indonesia in the PISA 2000 study). In addition, some of the patterns observed in the sections above for the PISA 2006 Science test seemed to be also present in the IEA/RLS and TIMSS results – for example, relatively lower correlations were observed in developing than in industrialized countries. Developing countries with non-Indo-European languages, like Botswana or Nigeria, had particularly low correlations in the IEA/RLS study.

Nobody will be surprised that in the IEA/RLS study, as well as in the three first PISA studies (and possibly in other international studies) the cognitive instruments were somewhat more appropriate, in cultural and linguistic terms, for the group of western countries that represented the majority of the participating countries.

The analyses in this paper suggest that the data from some other groups of countries might have suffered from a lesser level of equivalence. The problem seemed to affect, in particular, a number of Middle East and of Asian countries, and it might be really serious for possible comparisons that those countries would want to conduct within their own linguistic or geographic group. Suppose for example that a comparison is conducted between Tunisia and the other three Arabic-speaking countries that participated in PISA 2006, using the Science data described in this article. The average amount of bias affecting the results of any one of these countries would be no less than 33 per cent of the total variance in item difficulty. If the comparison involved Japan, Korea, Hong Kong and Taipei, the average amount of bias would be 30 per cent. By contrast, in case the same kind of comparison was conducted between Germany and the three other German-speaking countries participating in PISA 2006, the average amount of bias would be only 11 per cent.

There is a clear indication in these results that more in-depth analyses are needed, particularly for the Arabic and Chinese versions, to ascertain to what extent the problem could be caused by translation and adaptation. Some of the verifiers suggested that a special version of the translation guidelines should be prepared to help with the specificities of non Indo-European languages. The PISA international translation team could conduct a judgmental analysis of the items with DIF identified in those versions, with help from a panel of native translators, using a design similar to that employed by Ercikan and al. (2004) in their study of the English and French versions of the Canadian SAIP study.

Finally, the PISA Technical Advisory group will have to discuss whether a standard should be established (on the basis of the indicator proposed in this study, or on other indicators to be developed), to identify those countries where the version(s) of the test instruments in national language appear not to measure the latent construct in a way that is equivalent enough with the measures obtained in other countries to support meaningful international comparisons.

## References


Baye, A. (2004). La gestion des spécificités linguistiques et culturelles dans les évaluations internationales de la lecture, *Politiques d'éducation et de formation*, *11*, 55-70.

Blum, A., Goldstein, H., and Guerin-Pace, F. (2001). An analysis of international comparisons of adult literacy. *Assessment in Education*, *8*(2), 225-246.

Elley, W. B. (1992). *How in the world do students read? IEA study of reading literacy*. The Hague: IEA.

Ercikan, K., Gierl, M. J., McCreith, K, Puhan, G., Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17*(3), 301-321.

Ercikan, K., and Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS, *International Journal of Testing*, *5*(1), 23-35.

Grisay, A. (2003a). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, *20*(2), 225-240.

Grisay, A. (2003b). *PISA 2000: Differences in item difficulty between English, French and German countries,* Unpublished PISA TAG paper.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests, *European Journal of Psychological Assessment*, *10*, 229-240.

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

Le, L. (2006, April). *Analysis of differential item functioning in PISA 2006*. Paper prepared for the annual meeting of the American Educational Research Association, San Francisco.

McQueen, J., and Mendelovits, J. (2003). Cultural equivalence in a cross-cultural study. *Language Testing*, *20*(2), 208-225.

OECD (2007). *Technical standards for PISA 2009.* OECD document [GB(2007)4/REV1]. Retrieved from OECD, Directorate for Education web site: http://www.pisa.oecd.org

Van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence, *ZUMA-Nachrichten Spezial, 3*, 41-52.

Zumbo, B. D. (2003). Does item level DIF manifest itself in scale level analyses? Implications for translating language tests. *Language testing 20*(2), 136-147.