Provided for non-commercial research and educational use only.
Not for reproduction or distribution or commercial use.

Volume 33 Number 1 March 2007 ISSN 0191-491X

Studies in Educational Evaluation

Editor-in-Chief

David Nevo
School of Education
Tel-Aviv University, Israel

Associate Editors

Marvin C. Alkin
Department of Education
UCLA
Los Angeles
CA 90095
U.S.A.

Claus Carstensen
Institute for Science Education (IPN)
at the University of Kiel
Germany

This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

http://www.elsevier.com/locate/permissionusematerial

ELSEVIER

Studies in Educational Evaluation 33 (2007) 69–86

Studies in Educational Evaluation

www.elsevier.com/stueduc

MEASURING THE EQUIVALENCE OF ITEM DIFFICULTY IN THE VARIOUS VERSIONS OF AN INTERNATIONAL TEST

Aletta Grisay and Christian Monseur

University of Liège, Belgium

This article focuses on statistical analyses using PISA data to check for translation effects. This is an important article and one of which Arieh would have approved.

Abstract

In this article, data from the Reading Literacy study conducted in 2000 and 2001 as part of the OECD *Programme for International Student Assessment* (PISA) were analysed in order to explore equivalence issues across the 47 test adaptations in various languages of instruction that were used by the participating countries. On average, about 82% of the variance in relative item difficulty were found to be common across the various national versions. However, the index of equivalence appeared to be lower than desirable in certain categories of countries. Tentative analyses were conducted to better understand the reasons behind these differences.

International surveys of school systems require that all the dimensions of interest (e.g., students' proficiency in reading literacy; students' self-concept in mathematics; teachers' professional satisfaction, or schools' educational resources) have the same meaning across all participating countries in order to ensure valid comparisons. A preliminary condition to this requirement is the equivalence of the instruments in all languages of instruction used in the various school systems.

Equivalence of test items can be compromised when translation flaws or differences in students' familiarity with the question content or context cause item/country interactions (i.e., the item appears as easier or harder for students at a same level of proficiency in different countries). When too many cases of differential item functioning (DIF) occur in

0191-491X/$ – see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.stueduc.2007.01.006

an international test, the interpretation of the international scale becomes problematic, because the instability across countries of item difficulties prevents any accurate description of the skills associated with different scale points. In early IEA studies such as the Reading Comprehension study (Thorndike, 1973) possible differences in item functioning were identified by patiently checking cross-country similarities or dissimilarities in the item discrimination indices and in the patterns of attractiveness of the various answers in MC questions. Commonly used methods for detecting DIF items today include analyses based on Item Response Theory (IRT), logistic regressions controlling the difficulty of the items on the basis of observed scores, and the use of Mantel-Haenszel tests.

In this article, data from the reading literacy study conducted in 2000 and 2001 as part of the OECD *Programme for International Student Assessment* (PISA) are analysed in order to explore item-level equivalence issues across the various languages used in the participating countries.

The PISA studies are mainly intended to provide the OECD member countries (as well as a number of non-OECD "partner" countries) with comparative information on reading, mathematics and science literacy of their 15 years old students. The assessments are conducted on a 3 x 3-years periodical basis. Every 3 years one of the domains is assessed as the major domain and the two others as minor domains, so that regular information is collected on all three domains every 3 years and an in-depth study is conducted consecutively on each domain every 9 years. Items are field tested in participating countries on samples of approximately 1500 students. For the main study, samples of 4500 students per country are usually drawn.

Reading literacy was the major domain in the 2000 study, which was conducted in 28 OECD countries (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, Spain, Sweden, Switzerland, the United Kingdom and the USA) and in 11 partner countries (Brazil, Latvia and the Russian Federation participated in the study in 2000, and Albania, Argentina, Bulgaria, Chile, Hong Kong, Indonesia, Israel, Macedonia, Peru, Romania and Thailand joined a second round in 2001).

The reading test materials used in the PISA 2000 Main Study included 37 reading passages, each followed by a small set of paper and pencil questions, for a total of 132 items. These reading tasks were distributed into clusters each intended to last approximately thirty minutes of administration time. The clusters were rotated according to a matrix design (together with the mathematics and science clusters) across nine booklets composed of four clusters each. Each student participated in a 2-hours test session (about 50 to 60 test items).

In PISA studies, participating countries routinely receive the source materials from the International Centre in two versions (English and French), together with detailed translation guidelines (see for example OECD/PISA, 2004). They are requested to translate all test and questionnaire materials into their national language(s) using double forward translation by two independent translators and reconciliation by a third experienced party. As far as possible, it is recommended that the English source version be used for one of the independent translations, and the French source version for the second one. All reconciled

national versions are then submitted to the PISA International Centre for central verification of equivalence against the source versions by a pool of translators specially appointed and trained.

Using materials in two languages as source versions was aimed at reducing to some extent the impact of linguistic and cultural specificities of the language used at the test development phase (which was English, even for those reading tasks that were derived from materials submitted by the participating countries in other national languages). To ensure equivalence between the two source versions, each test unit was double-translated and reconciled into French at an early stage during the development of the English source version, so that information from the French translators and reconcilers about possible difficulties or translation traps could be fed back to the test developers. In a surprisingly large number of cases this resulted in improvements being made to both the English and French sources, in order to make them as equivalent as possible and safer to translate into other PISA languages.[1]

Amount of Differential Item Functioning Within and Between Language Groups

If perfectly equivalent translations of the PISA 2000 reading materials had been obtained through these procedures in all participating countries, this would mean that no language-related DIF would be observed in the data. In other words, the differences in relative item difficulty would not be larger when comparing, for example, the item parameters from a French-, German-, or Portuguese-speaking country with those from the English-speaking countries than when comparing the item parameters within the group of English-speaking countries themselves. Item/country interactions can exist within groups of countries sharing a same language, but of course all or almost all of them would be due to cultural or curricular specificities rather than to linguistic differences. If the number of DIF items appears to be significantly larger when comparing versions translated into different languages than when comparing homolingual versions, then the difference can inform us about the average magnitude of the measurement bias introduced by translation.

Item difficulty parameters (centred deltas) were obtained by running separate Rasch analyses on the country datasets corresponding to the 47 language versions used by substantial proportions of students in each national sample.[2] These included:

- 7 English national versions used in English-speaking countries (directly adapted from the English source version with only very minor changes);
- 4 French national versions used in French-speaking countries (directly adapted from the French source version with only very minor changes);
- 4 German national versions (adapted with only very minor changes from a common German version developed co-operatively by the four German-speaking countries and carefully cross-checked against each other and against the two source versions);
- 5 Spanish national versions, most of which were developed independently by the various Spanish-speaking countries, with no or limited communication between them nor systematic sharing of the translation work.

Table 1: PISA 2000 Reading: Percent of Items with DIF of 0.5 or More by Reference to the Average Item Difficulties Observed in …[3]

	…the group of English versions	…the group of French versions	…the group of German versions	…the group of Spanish versions	…all participating countries
English versions					
IRL.ENG	4.8%	28.0%	28.0%	31.2%	14.4%
CAN.ENG	0.0%	27.2%	30.4%	26.4%	15.2%
AUS.ENG	0.0%	27.2%	23.2%	28.8%	16.0%
NZL.ENG	0.8%	28.0%	24.8%	32.8%	16.8%
QSC.ENG	1.6%	29.6%	31.2%	32.0%	19.2%
QUK.ENG	3.2%	29.6%	25.6%	34.4%	20.8%
USA.ENG	6.4%	36.0%	35.2%	33.6%	22.4%
Mean ENG	2.4%	29.3%	28.3%	31.3%	17.8%
French versions					
QBR.FRE	24.8%	0.0%	25.6%	23.2%	16.0%
CHE.FRE	32.0%	2.4%	24.0%	28.8%	17.6%
CAN.FRE	21.6%	5.6%	21.6%	28.8%	18.4%
FRA.FRE	36.0%	2.4%	31.2%	31.2%	24.0%
Mean FRE	28.6%	2.6%	25.6%	28.0%	19.0%
German versions					
LUX.GER	22.4%	19.2%	1.6%	24.8%	15.2%
DEU.GER	25.6%	22.4%	0.0%	31.2%	16.0%
CHE.GER	27.2%	23.2%	5.6%	33.6%	18.4%
AUT.GER	28.8%	29.6%	1.6%	37.6%	20.0%
Mean GER	26.0%	23.6%	2.2%	31.8%	17.4%
Spanish versions					
ESP.SPA	28.8%	22.4%	27.2%	17.6%	17.6%
ARG.SPA	28.2%	25.8%	36.3%	9.7%	20.2%
CHL.SPA	38.7%	40.3%	41.9%	12.1%	29.0%
MEX.SPA	44.0%	44.8%	46.4%	22.4%	32.8%
PER.SPA	53.2%	47.6%	46.8%	23.4%	38.7%
Mean SPA	38.6%	36.2%	39.7%	17.0%	27.6%
Asian versions					
HGK.CHI	36.8%	44.8%	43.2%	41.6%	32.8%
JPN.JAP	43.2%	38.4%	36.8%	45.6%	35.2%
THA.THA	48.4%	46.0%	47.6%	37.9%	36.3%
KOR.KOR	45.6%	45.6%	45.6%	53.6%	36.8%
IDN.IND	48.4%	55.6%	50.0%	45.2%	42.7%
Mean Asian	44.5%	46.1%	44.6%	44.8%	36.7%

These 20 national versions were used to assess the amount of DIF items within and between language groups. Average item deltas were computed for each of the four languages as well as for the whole international sample (all 47 versions). Table 1 contains, for each version in each of the four language groups, the percentages of items where the national delta differed by more than 0.5 logits from the average item difficulties observed (i) in the same language group (grey diagonal cells); (ii) in the three other language groups, and (3) at the international level.

The grey diagonal in Table 1 is of particular interest. As expected, the three groups of countries that used an *identical* version of the materials, except for a very limited number of adaptations, had extremely low percentages of items with DIF *within their own English, French or German language group* (in each group less than 3% on average, with a maximum of 5 or 6% for national versions that had the largest number of adaptations). By contrast, the number of within-language DIFs was considerably higher in the group of Spanish countries, where independent or almost independent translations were used (17% of DIF items in average, ranging from 10% to 23%).

Clearly, in order to minimize item/country interactions in countries sharing the same language of instruction, using that *same language* is not sufficient. Strictly *identical versions* are a much better solution, which gives some idea of the net impact of the translation factor. Any translated version is just one in an infinite number of potential "sister" translations that might be derived from a same source text in a given target language, and the Spanish example indicates that any two of these "sister versions" are likely to behave in a less equivalent way than authentic "twin versions".

The white cells in Table 1 present the proportion of DIF items when comparing "cousin versions", that is, versions in languages like English, French, German and Spanish, that are different but can all be classified in the same Indo-European language family. As expected, the average percent of DIF items is higher again (25 to 30%) in comparisons involving "cousin" languages than in the two previous same-language cases. This includes the comparisons between the relative item difficulties in countries using the two 'parallel' English *vs.* French source versions.

Finally, in the dark grey section at the bottom of Table 1 it can be seen that the proportion of DIF items is maximal (about 45% on average) when comparing the relative item difficulties between the English, French, German and Spanish language groups and the versions used in a few Asian countries, i.e. in non-Indo-European languages that may be considered as the most linguistically and culturally "distant" among the languages used in the participating countries.

Amount of Between-Countries Variance in Item Difficulty Attributable to DIF

The data in Table 1 provide rather clear evidence that part of the equivalence of a test instrument is *always* lost when translating it into other languages. It can also be seen that, while the amount of DIF items in the PISA 2000 Reading test varied considerably, depending upon the "proximity" of the versions and languages included in the comparisons, it was never negligible in all cases when the languages differed: even when comparing the materials directly adapted from the English source version to those adapted from French, the proportion of DIF items was consistently around 25-30%.

These figures are not a surprise. Even higher proportions of DIF items were identified by Ercikan (2005) in a study where she compared the versions of the TIMSS 1995 mathematics and science tests used in England, France, the United States and Canada (English- and French-speaking provinces). Both in mathematics and science, the least differences in item relative difficulties were observed between the two Canadian versions (14% of DIF items in mathematics and 37% in science). The largest ones were observed between the versions used in the United States and France (59% in mathematics and 79% in science).

However, the *meaning* of this kind of finding (i.e., the exact extent to which the amount of DIF identified in the various versions threatens the comparability of the TIMSS 1995 or of the PISA 2000 results) is not easily determined by simply counting the number of DIF items. This is partly because the definition of a DIF item is somewhat dependent upon the method used to identify them, and partly because the actual magnitude of DIFs can vary a lot beyond the threshold of 0.5 logits, the value usually considered as indicating "moderate" DIF.

In order to estimate the *total variance in item difficulties that is common across the various versions of an international test*, Grisay, de Jong, Gebhardt, Berezner, and Halleux (2006) proposed retaining the communalities from a factor analysis where the item difficulties were used as observations and the test versions as variables. They calculated this indicator of equivalence in relative item difficulty for the science test in the PISA 2006 Field Trial. The indicator proved to have interesting characteristics – in particular, it could be seen that there was a significant gap between the versions in Indo-European and most of those in *non* Indo-European languages. The latter group had much lower commonalities than the former.

The commonalities from the same type of factor analysis, obtained using the PISA 2000 Reading data (item deltas for 124 reading items in 47 national versions) have been presented in Figure 1.

The common factor extracted by the analysis explained 82% of the total variance; all versions except 13 had loadings of more than 0.90 on that factor, and for only one version the loading was less than 0.80, indicating substantial comparability of the relative item difficulties across all countries.[4] The commonalities appeared to be high and reasonably consistent across the majority of test versions used in western countries. Interestingly, neither the seven English versions nor any other language group formed a "dominant" cluster at the top of the graph.

By contrast, the group of versions that were located at the bottom of the graphic because of their lower communalities (less than 80%) seemed to be composed of two specific categories, including all versions in Asian languages and most of the versions from participating countries with low GDP and low reading scores in the PISA assessment.

The unique component of variance that remained unexplained for each national version after extracting the common factor can be considered as an accurate indicator of the total amount of DIF contained in that version when comparing its relative item difficulties with those observed, in average, in the whole set of participating countries. In fact, the correlation between communalities and proportion of items with DIF more than 0.5 logits was - 0.88 for $N = 47$ versions.

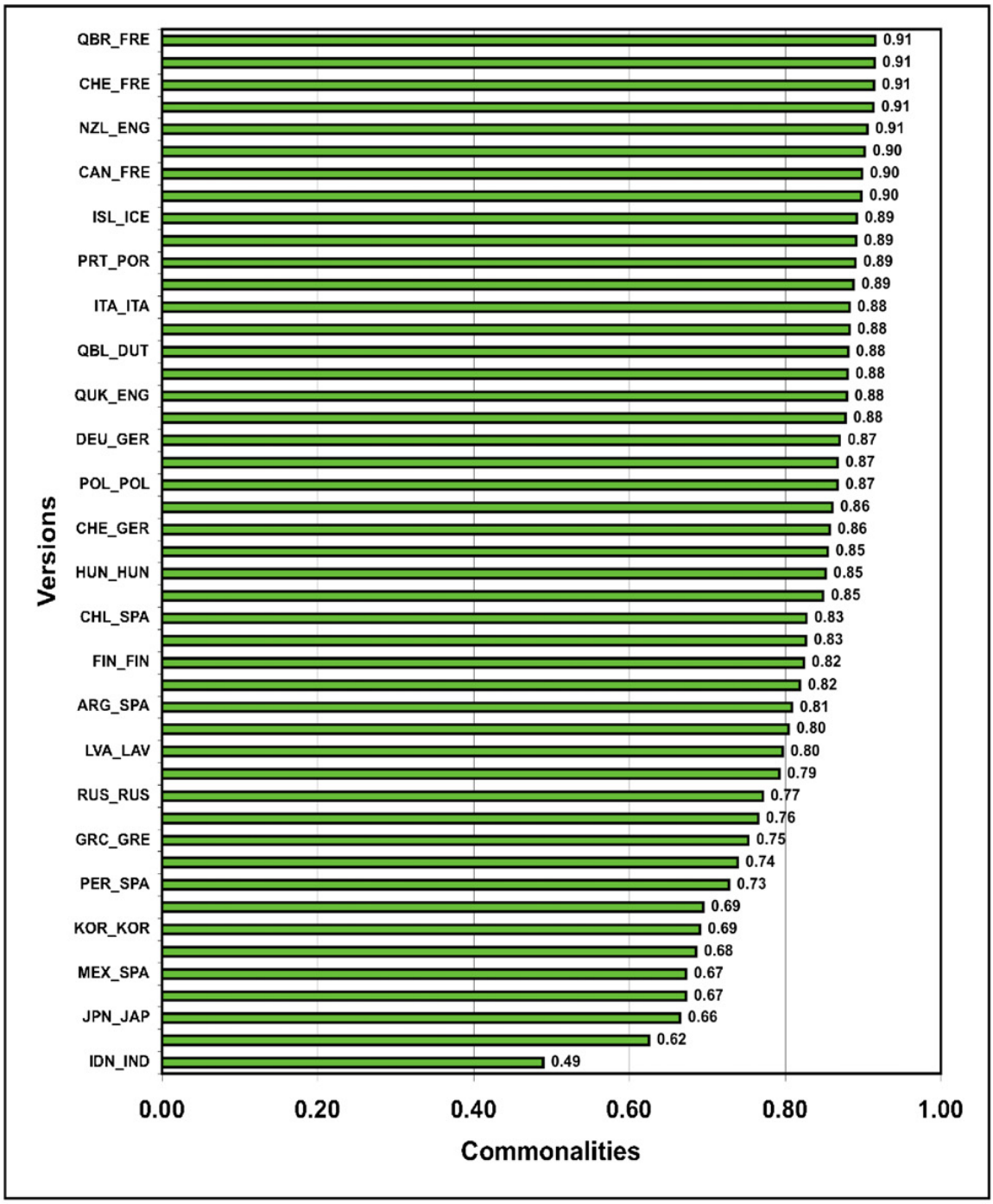

Figure 1: PISA 2000 Reading: Commonalities in Item Difficulty Across National Versions

Some Correlates of the Commonality Indicator

To explore further the pattern in Figure 1, the correlations observed between the indicator of commonality and a few characteristics of the participating countries have been presented in Table 2: country GDP per capita, expressed in US dollars at purchasing power

parity (PPP); country mean PISA score in reading; reliability of the national version of the test;[5] and two dichotomised variables (Indo-European or non Indo-European language; Asian or non-Asian country).

Table 2: Correlations Between Commonality and Selected Characteristics of PISA Countries

$N = 47$ versions	Commonality	GDP per capita	Country mean Reading score	Reliability	Indo-European language
Commonality in item difficulties	1.				
GDP per capita	0.544**	1.			
Country mean reading score	0.443*	0.627**	1.		
Reliability	0.723**	0.345*	0.147	1.	
Indo-European language	0.559**	0.284	0.072	0.518*	1.
Asian country	- 0.705**	- 0.145	0.134	- 0.707**	- 0.624*

** Significant for $p < 0001$
* Significant for $p < 05$

It must be noted that some of these characteristics are interrelated. For example, out of the eight versions in non Indo-European languages, five were used in Asian countries and only three (Finnish, Hungarian and Hebrew) in other geographic regions, thus the variables *Indo-European language* and *Asian country* partly conveyed "mirror" information. Similarly, developing countries tend to have lower PISA scores than the industrialised countries, and thus it was no surprise that *GDP per capita* had a strong negative correlation with *Mean PISA score*.

The very high negative correlation between *commonality in item difficulties* and *Asian country* (-0.705) confirmed that a significantly lower level of psychometric equivalence was obtained in the group of countries that were the most linguistically and culturally "distant" from the majority of the OECD member countries.

More surprising was the highly positive correlation observed between *commonality* and *test reliability* (0.723), as well as the negative correlation between *reliability* and *Asian country* (-0.707).

In theory, the proportion of DIF items does not necessarily affect the reliability estimates. Zumbo (2003) used a simulation technique to investigate to what extent the indicators commonly used to compare scale-level equivalence of constructs across the various versions of an international test, such as Cronbach's alpha, or RMSEA fit statistics (root mean square error of approximations) from multi-group confirmatory factor analyses (CFA) were affected by different proportions of items with DIF, purposefully included in a fictional dataset. He observed that alphas and RMSEAs were only very marginally affected in the simulation when the proportion of items with DIF was decreased to less than 3% or increased to more than 40% of the items. The relationship observed in the PISA Reading

test between communalities and reliabilities suggested therefore that some *non-random* factor affecting the geographic or cultural distribution of DIF items was deteriorating, to some extent, the reliability of the scale in a number of countries. Two potential explanations are offered in the last two sections of this article.

Possible Differential Effect of Item Format in Certain Groups of Countries

The PISA Reading test included about 45% multiple-choice items and 55% constructed response items. A first analysis was thus conducted to explore potential systematic biases due to item format. Multiple choice items might tend to function differently in certain groups of countries, either because the MC format is less familiar than other formats in those school systems, or because the translation of MC items can be problematic in languages where order of words and syntactic dependencies are very different from the English and French source versions. On the other hand, constructed response items could suffer from lower discrimination indices in developing countries with disproportionate numbers of low-achieving students, due to their low performances in writing.

Some evidence confirming this hypothesis may be found in Figure 2, where the commonalities have been presented that would be obtained for each national version if the PISA 2000 reading test was *only* composed of multiple-choice items (MC) or, alternatively, *only* of constructed response items (CR).

With few exceptions, the versions used in European and North American countries would have had slightly higher commonalities if the test had been composed exclusively of CR items than if it had been a multiple-choice test – a difference that may be due to the guessing component usually observed for multiple-choice instruments. However, two contrasting phenomena were observed near and at the bottom of the graph:

(i) On one hand, a number of countries with low GDP and average country scores well below the OECD mean proficiency level had MC commonalities significantly higher than their CR commonalities (Argentina, Bulgaria, Macedonia, Albania, Indonesia, and, most of all, Brazil and Mexico), probably confirming that constructed response items tend to be less stable than MC items in low-performing countries. The proportion of items with DIF > 0.5 was indeed higher for CR than for MC items in those countries (30% *vs* 26%).

(ii) On the other hand, in four out of five Asian countries (Korea, Hong Kong, Japan, and to a lesser extent Thailand), the CR commonalities were far higher than the MC commonalities, probably confirming that many multiple-choice items do not function in equivalent ways in their versions, compared to the test versions in Western languages. In this group of countries, the proportion of items with DIF > 0.5 was 40% for MC items *vs.* 35% for the CR items.

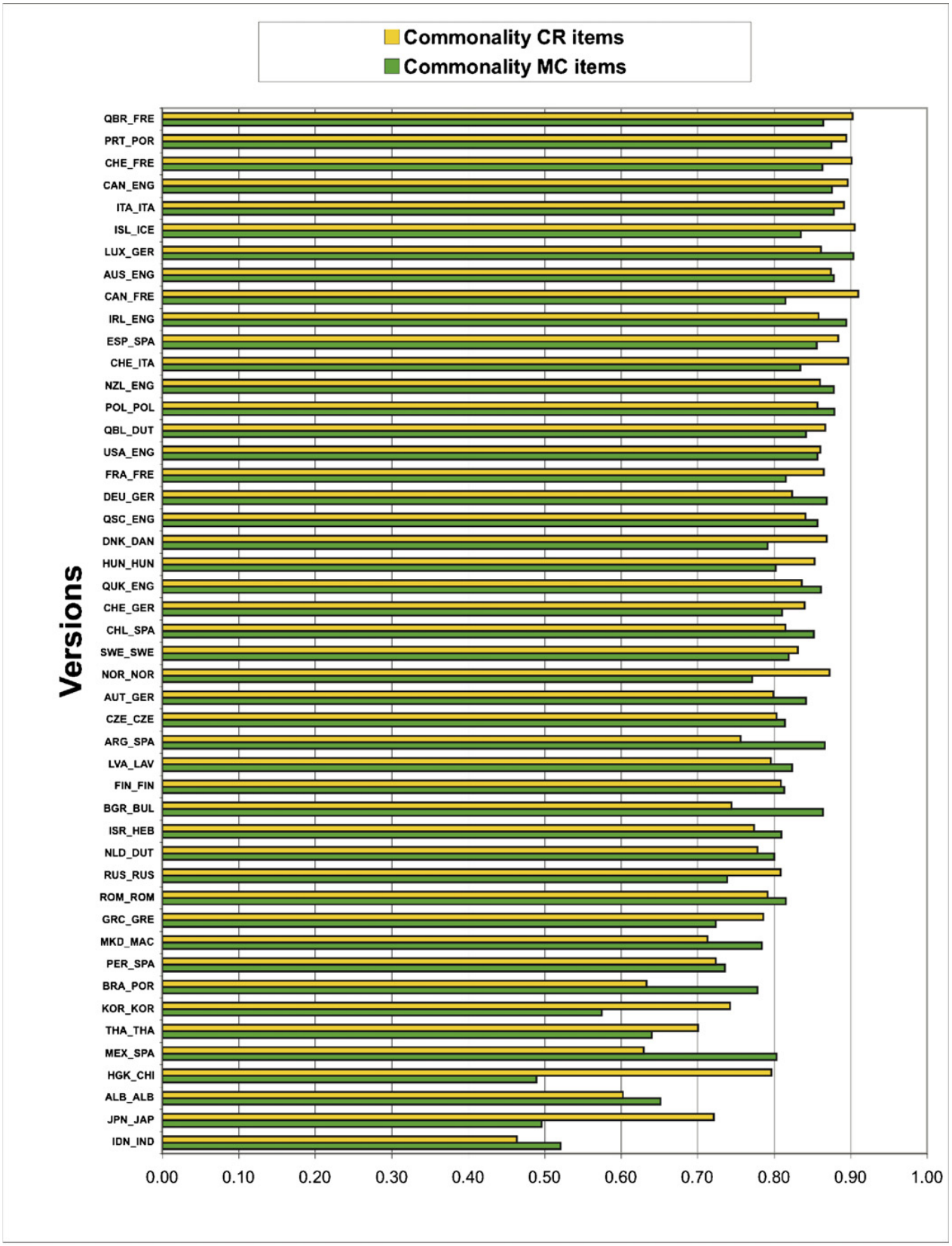

Figure 2: PISA 2000 Reading: Commonalities for Potential Tests Composed Only of Multiple Choice or Constructed Response Items

Possible Effect Due to Local Dependency of Items Related to the Same Reading Passage[6]

As mentioned above, the PISA 2000 Reading items were embedded into 37 units. Each unit consisted of a text followed by a few questions (usually 3-4). Therefore all items within a unit were interrelated as they were all based on the same stimulus. Ignoring the potential effects of this item clustering, which violates the IRT assumption of item independency, could result in some underestimation of the measurement error and overestimation of the scale reliability, proportional to the magnitude of item dependency.

In an international assessment, item dependency might also vary from country to country, for at least two reasons:

1. A cultural issue: teenagers in certain countries might be more familiar than in other countries with the topic discussed in the stimulus.
2. The difficulty level of the stimulus: if the text is too difficult to read, low achievers might simply skip all attached questions (including the easy ones) because they don't understand it. Countries with large proportions of low-achieving students would then suffer from larger item dependency effects.

Several procedures can be used to detect local item dependency (LID) (See Gao & Wang, 2005; Zenisky, Hambleton, & Sireci, 2006). In particular, according to Zenisky et al.

> LID on a test can be ascertained by comparing two separate reliability estimates. The first estimate assumes that all items are locally independent and ignores the testlet structure. The second estimate models the inherent testlet structure, which involves forming testlets for all context dependent item sets. If the testlet-based reliability estimate is substantially lower than the item-based estimate, LID is present. (2006, pp. 7-8)

The PISA 2000 cognitive data were recoded into a rotated test design of 37 units and calibrated using partial credit model. With a rotated design, it is inadequate to compute the Crombach ! reliability coefficient, due to the amount of missing data. Therefore Crombach ! coefficients were computed *per booklet* for two models:

1. The model used to originally scale the PISA 2000 data, which considered all items to be independent;

2. Partial credit model, where each unit was considered as a single polytomous item, with multiple score levels.

Table 3 presents the decrease of reliability per booklet and per country (with a decomposition per language where feasible) between the two models.

As shown in Table 3, on average, there was a small decrease in reliability coefficients, which indicated some LID in the data. However, the decrease varied somewhat from country to country. In particular, the English-speaking countries had the

smallest average decrease in reliability, while the two Asian countries included in this analysis, i.e. Japan and Korea, were among the countries with the highest decreases.

Table 3: PISA 2000 Reading: Reduction of Reliability per Country per Booklet Due to Local Item Dependency

	Countries are ordered by increasing mean values									
	B1	B2	B3	B4	B5	B6	B7	B8	B9	Mean
NLD_DUT	0.03	0.04	0.03	0.02	0.02	0.04	0.03	0.03	0.03	0.03
QSC_ENG	0.02	0.06	0.04	0.02	0.02	0.03	0.04	0.05	0.03	0.03
NZL_ENG	0.03	0.06	0.04	0.03	0.03	0.03	0.04	0.04	0.03	0.04
QUK_ENG	0.03	0.07	0.03	0.03	0.02	0.03	0.04	0.04	0.03	0.04
USA_ENG	0.03	0.07	0.04	0.03	0.02	0.03	0.04	0.04	0.03	0.04
IRL_ENG	0.03	0.07	0.04	0.03	0.03	0.03	0.04	0.05	0.03	0.04
AUS_ENG	0.03	0.07	0.04	0.03	0.02	0.03	0.04	0.05	0.04	0.04
QBL_DUT	0.03	0.06	0.05	0.04	0.02	0.04	0.05	0.04	0.04	0.04
PRT_POR	0.03	0.06	0.04	0.04	0.03	0.04	0.04	0.05	0.03	0.04
MEX_SPA	0.03	0.06	0.04	0.04	0.04	0.04	0.04	0.05	0.04	0.04
CAN_ENG/FRE	0.03	0.07	0.04	0.03	0.03	0.04	0.04	0.05	0.04	0.04
AUT_GER	0.03	0.09	0.04	0.04	0.03	0.04	0.04	0.05	0.04	0.04
NOR_NOR	0.04	0.07	0.04	0.03	0.03	0.04	0.04	0.04	0.05	0.04
FIN_FIN	0.03	0.06	0.04	0.04	0.03	0.05	0.04	0.05	0.05	0.04
SWE_SWE	0.03	0.08	0.03	0.04	0.03	0.04	0.05	0.07	0.04	0.04
QBR_FRE	0.03	0.07	0.05	0.03	0.03	0.04	0.04	0.06	0.05	0.04
LVA_LAV	0.03	0.03	0.04	0.05	0.04	0.05	0.05	0.06	0.04	0.04
POL_POL	0.03	0.07	0.05	0.04	0.04	0.03	0.04	0.06	0.04	0.04
HUN_HUN	0.03	0.03	0.06	0.04	0.05	0.05	0.05	0.05	0.04	0.04
LUX_GER	0.03	0.08	0.04	0.04	0.02	0.04	0.05	0.06	0.04	0.05
ISL_ICE	0.03	0.06	0.05	0.03	0.02	0.04	0.06	0.07	0.04	0.05
FRA_FRE	0.03	0.08	0.02	0.04	0.03	0.06	0.05	0.07	0.04	0.05
DEU_GER	0.02	0.09	0.04	0.04	0.03	0.05	0.05	0.05	0.04	0.05
CHD_GER	0.03	0.10	0.04	0.05	0.03	0.04	0.04	0.05	0.04	0.05
CZE_CZE	0.03	0.07	0.05	0.05	0.04	0.04	0.05	0.06	0.05	0.05
DNK_DAN	0.03	0.08	0.04	0.04	0.05	0.05	0.05	0.06	0.04	0.05
GRC_GRE	0.04	0.08	0.05	0.05	0.04	0.04	0.04	0.06	0.06	0.05
CHE_FRE	0.03	0.08	0.05	0.04	0.03	0.04	0.05	0.07	0.06	0.05
SPA_SPA	0.03	0.08	0.05	0.04	0.03	0.04	0.06	0.08	0.04	0.05
JPN_JAP	0.03	0.07	0.05	0.05	0.04	0.05	0.06	0.07	0.04	0.05
KOR_KOR	0.04	0.08	0.05	0.05	0.03	0.04	0.05	0.07	0.07	0.05
BRA_POR	0.04	0.08	0.05	0.06	0.05	0.05	0.05	0.06	0.05	0.05
ITA_ITA	0.04	0.08	0.06	0.05	0.05	0.05	0.06	0.08	0.05	0.06
RUS_RUS	0.04	0.11	0.04	0.05	0.05	0.06	0.06	0.07	0.05	0.06
CHE_ITA	0.05	0.08	0.06	0.07	0.05	0.06	0.09	0.11	0.08	0.07

Further, as shown in Figure 3, low achieving countries seemed to have, on average, slightly higher decreases in the ! coefficient. The correlation between decrease in reliability and country score was equal to -0.45 (-0.51 when excluding Mexico). In other words, low-achieving countries tended to have slightly higher local dependency.

There was a modest correlation between decrease in reliability and country communalities (-0.20, or -0.33 when excluding the Italian version used in Switzerland), suggesting that the communalities of low achieving countries might have been slightly affected by their higher local item dependency.

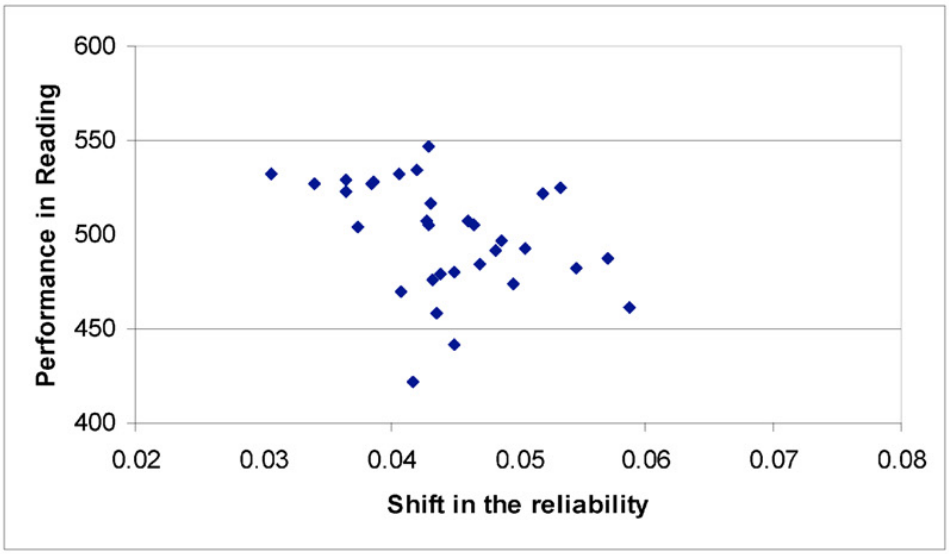

Figure 3: PISA 2000 Reading: Relationship Between Country Performance in Reading and Decrease in Reliability Due to Local Item Dependency

It is worth noting that local dependency may also vary from unit to unit. To identify units with higher local dependency, the average covariance between items within units was computed by comparing the unit variance and the sum of the item variances included in that unit. This difference was then divided by two times the number of item covariance coefficients for that unit. Out of the 37 units, three units presented an average covariance higher than 0.10, i.e. R067 *Aesop* (0.18), R111 *Exchange* (0.12) and R122 *Just Art* (0.11). All three units had been identified as culturally sensitive during the PISA scaling procedures, and rather high rates of item/country interactions had been observed for several of their items.

Conclusion

The main results from the various analyses described above convey mixed messages about the equivalence of the test instruments used in the PISA 2000 assessment of reading literacy.

On the one hand, an indicator of overall commonality in item difficulties was developed, which appeared to have reasonably high values in most of the participating

countries. Reading literacy is usually viewed as a domain much more dependent than mathematics upon the linguistic and cultural characteristics of the assessed populations. This appeared to be true, but only to a moderate extent: the average commonality of item difficulties was as high as 91% in the PISA 2003 mathematics assessment (Grisay et al., 2006), but it still was no less than 82% in the PISA 2000 Reading study.

On the other hand, some disturbing findings also came to the fore, indicating that the variance in item difficulty that was *not* explained by this common factor could not be considered as random variance equally distributed across all participating countries, and deserved careful exploration.

First, when comparing the proportions of items with differential item functioning (DIF) across homolingual and heterolingual groups of countries, the analysis indicated that translating a test from a source version *had always at least a basic cost in terms of loss of equivalence*, whatever the quality of the translation. As shown in Table 1, even in comparisons where the two versions were as parallel as possible and any cultural and geographic difference was minimized (such as the comparisons involving the English- *vs* French-speaking provinces in Canada or the German- *vs* the French-speaking cantons in Switzerland), the number of DIF items was much higher than when comparing same-language versions in countries as far from each other as Ireland, New Zealand and the USA.

Second, the values of the indicator of commonality in item difficulty appeared to be lower than desirable for two groups of countries:

- All five Asian countries participating in PISA 2000, some of which were among the best-performing participants (Japan, Korea, and Hong-Kong);
- A number of low-GDP countries, also characterised by low average achievement in reading of their students. This group included one OECD country (Mexico), and a number of non-OECD "partner" countries participating in PISA (Albania, Bulgaria, Macedonia, Brazil, Peru, Indonesia and Thailand).

It is hardly surprising that equivalence in item difficulty was lower in these two groups of countries, which are most "distant" from the majority of OECD countries in terms of language or/and in terms of economic and societal characteristics. However, the exact psychometric mechanisms that play a role in these differences proved to be too complex and difficult to disentangle from each other, because of their partly contradictory effects.

In this article, two of these potential mechanisms were explored: (i) cross country differences in the students' reactions to multiple choice *vs* open-ended questions, and (ii) higher effects of item clustering into units in countries with high proportions of low-achieving students.

The differential impact of item format was confirmed, to a large extent, when computing the commonalities in separate analyses using either multiple-choice or constructed-response items. In four out of five Asian countries, the commonality obtained using multiple-choice items was far lower than that obtained for constructed-response items, possibly indicating that the MC format is more sensitive to large linguistic differences affecting syntax, order of sentence, or direction of writing. The reverse was true

for a number of countries with low GDP and low reading scores, where a reading test composed of MC items only would have had a much better commonality, possibly suggesting that in those countries the functioning of constructed-response items is partly biased by a writing skills factor.

Some impact was also confirmed (although it was rather modest) as regards the cross-country variation in local item dependency due to item clustering into test units. In all countries, when the data were rescaled using a partial credit model taking into account the unit design, a small decrease in alpha reliabilities was observed – as expected when local item dependency is present. This effect appeared to be slightly higher in low-performing countries, but also, interestingly, in the two high-performing Asian countries included in the analysis, Japan and Korea.

Some recommendations can be derived from these findings for the improvement of equivalence in future studies.

First, the development of a single "common" version should be strongly encouraged for all participating countries that share a same language of instruction. Independent translation is often preferred by countries that consider their national language to be "too different" from idiolects of the same language that are spoken in other countries. The evidence from the study suggests that in fact, using a same version (with adaptations to cope with national specificities) results in far higher levels of equivalence than independent versions. Particular attention should be paid, in the PISA 2009 study (where reading will again be the major domain) to the Spanish, Arabic and Chinese versions, for which no common version could be developed in the previous PISA studies.

Second, further exploration of the potential interaction of item format (MC or CR) with the linguistic and cultural characteristics of certain groups of countries might be required. Judgemental reviews of the DIF items that seem most typical in specific versions could probably be used to improve the translation guidelines provided to the national translation teams.

Third, although the overall impact of item dependency due to the clustering of items into units proved to be only marginal, the PISA test developers should keep in mind this aspect of the test design, particularly at the Field Trial phase, where specific units may appear to have larger dependency effects than desirable, and therefore not to be suitable for international use.

Notes

1. See Grisay (2003) for more detailed information on the translation procedures used in PISA and on the development of the French source version.
2. Minority languages only used in a limited number of sampled schools, such as German in Belgium, Swedish in Finland, Basque in Spain etc., were omitted from the analysis.
   Some of the items were dropped or modified in the group of countries that joined the study in 2001. Our analyses are limited to the 124 items that received no modification and were available in all countries.

3. Here is a key for abbreviations used in this illustration and elsewhere in this article:

Abbreviation	Country or sub-national entity	Language
ALB_ALB	Albania	Albanian
ARG_SPA	Argentina	Spanish
AUS_ENG	Australia	English
AUT_GER	Austria	German
BRA_POR	Brazil	Portuguese
BGR_BUL	Bulgaria	Bulgar
CAN_ENG	Canada	English
CAN_FRE	Canada	French
CHE_GER	Switzerland	German
CHE_FRE	Switzerland	French
CHE_ITA	Switzerland	Italian
CHL_SPA	Chile	Spanish
CZE_CZE	Czech Republic	Czech
DEU_GER	Germany	German
DNK_DAN	Denmark	Danish
ESP_SPA	Spain	Spanish
FIN_FIN	Finland	Finnish
FRA_FRE	France	French
GRC_GRE	Greece	Greek
HGK_CHI	Hong Kong	Chinese
HUN_HUN	Hungary	Hungarian
IDN_IND	Indonesia	Bahasa Indonesian
IRL_ENG	Ireland	English
ISL_ICE	Iceland	Icelandic
ISR_HEB	Israel	Hebrew
ITA_ITA	Italy	Italian
JPN_JAP	Japan	Japanese
KOR_KOR	Korea (Republic of --)	Korean
LUX_GER	Luxembourg	German
LVA_LAV	Latvia	Latvian
MEX_SPA	Mexico	Spanish
MKD_MAC	FYR Macedonia	Macedonian
NLD_DUT	Netherlands	Dutch
NOR_NOR	Norway	Norwegian
NZL_ENG	New Zealand	English
PER_SPA	Peru	Spanish
POL_POL	Poland	Polish
PRT_POR	Portugal	Portuguese
QBL_DUT	Belgium	Dutch
QBR_FRE	Belgium	French
QSC_ENG	Scotland	English
QUK_ENG	England	English
ROM_ROM	Romania	Romanian
RUS_RUS	Russian Federation	Russian
SWE_SWE	Sweden	Swedish
THA_THA	Thailand	Thai
USA_ENG	United States	English

4. These results differ slightly from those obtained in the factor analysis conducted by Baye (2004) on the same set of reading data, but without including the 11 partner countries that joined the study in 2001. In her analysis, the G factor was the same (82% of the total variance and high loadings for much the same versions), but two very minor and un-interpretable other factors had Eigen values more than 1.
5. The proxy measure of reliability used here for each national version is the average correlation between the five plausible values in reading that were drawn for the students who were assessed using that version of the test.
6. The analyses in this section required complete rescaling of each of the datasets included. Because of time constraints and other practical reasons, they were conducted for a reduced set of 35 national versions, excluding most of the non-OECD "partner" countries.

References

Baye, A. (2004). La gestion des spécificités linguistiques et culturelles dans les évaluations internationales de la lecture [How to deal with linguistic and cultural specificities in international assessments of reading literacy]. *Politiques d'éducation et de formation*, *11*, 55-70.

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

Gao, L., & Wang, C. (2005), *Using five procedures to detect DIF with passage-based testlets*. Paper prepared for the annual meeting of the National Council of Measurement in Education, Montreal, Quebec.

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

Grisay, A., de Jong, J.H.L., Gebhardt, E., Berezner, A., & Halleux, B. (2006). *Translation equivalence across PISA countries*. Paper presented at the 5th Conference of the International Test Commission, Brussels, Belgium, 6-8 July 2006.

OECD/PISA (2004), *PISA 2006 translation and adaptation guidelines*, Doc. NPM(0409)13.

Thorndike, R.L. (1973). *Reading comprehension education in fifteen countries: An empirical study.* Stockholm: Almquist & Wiksell.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2006) *Effects of local dependence on the validity of IRT item, test and ability statistics*. Available at: http://www.aamc.org/students/mcat/research/

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.

The Authors

ALETTA GRISAY graduated in 1963 in Philosophy and Literature at the University of Liège, Belgium, where she spent spent most of her research career at the Service de Pédagogie Expérimentale created by Prof. G. De Landsheere. She was involved in a

number of IEA surveys (Reading Comprehension Study, Literature Study, English as Foreign Language Study, Reading Literacy Study and Civic Education Study). She is currently a member of the Technical Advisory Group of the OECD/PISA programme.

CHRISTIAN MONSEUR is professor at the University of Liège in Belgium. He is also an associate researcher for the Australian Council of Educational Research. Before assuming these roles, Christian was the data manager for PISA 2000 and director of the PISA Plus project. He has a qualification as a teacher, has graduated in Educational Sciences, has completed a Master's degree in statistics and received a Ph.D. in 2005. He has published a number of articles and chapters in the field of educational assessment.

Correspondence: <agrisay@attglobal.net>

Author's personal c