Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment.

Svend Kreiner
Dept. of Biostatistics, University of Copenhagen

Summary. This paper addresses methodological issues related to the scaling model underlying the international comparison of student attainment in the Programme for International Student Attainment (PISA) and, in particular, the question of whether PISA's ranking of countries is confounded by differential item functioning (DIF). To address these issues, we reanalyze data on reading from the 2006 survey. The analysis provides strong evidence that the fit of item responses to PISA's scaling model is inadequate and very strong evidence of DIF that affects the ranking of countries. The second part of the paper presents two ways to deal with the problems caused by the inadequate scaling model: 1) modelling departures from the scaling model so that measurement can be adjusted for DIF and other problems before countries are compared, and 2) purification by elimination of items that do not agree with the scaling model. Analyses of data on reading from UK and Denmark illustrate these methods. The attempt to model departures from the scaling model is successful for subsets of items relating to different reading units, but fails to identify an adequate model for the complete set of items. Purification is more successful pin-pointing a subset of eight items that fits PISA's scaling model in UK and Denmark.

Keywords: Differential item functioning; Educational testing; Programme for International Student Assessment; Rasch models; Reading literacy

1. Introduction

The Programme for International Student Assessment (PISA) represents an ambitious and expensive large-scale attempt to measure and compare literacy in reading, mathematics and science in a large number of countries. The first PISA survey was launched in 2000, and it has since been followed up with surveys in 2003, 2006 and 2009. Many concerns have been raised concerning the comparability of educational test results from different countries in general and in particular with the difficulties in producing items that are culturally and linguistically neutral. We refer to Prais (2003), Goldstein (2004), Brown et.al. (2007), and Hopmann, Brinek & Retzl (2007) who discuss many of these issues.

A closer look at the foundation of PISA reveals that PISA rests on the assumption that item responses fit a specific type of scaling model referred to as a Rasch (1960/1980) model. Some efforts have been invested in checking the fit of item responses to the model and in particular the assumption that item difficulties were homogenous across countries. Purification that eliminated items that did not fit the Rasch model was also carried out during PISA's initial item analysis, but the efforts appear to be perfunctory and unsystematic, and no attempts appear to have been made to check that the items that survived the process actually fit the scaling model.

PISA's comparison of countries relies on so-called plausible student scores derived from the Rasch model. For this reason, the lack of published evidence supporting the scaling model used by PISA and in particular the suspicion that some items may function differently in different countries raise serious questions concerning the credibility of PISA's ranking of countries. We address these issues by reanalyses of data on reading from PISA 2006, where we check the adequacy of the Rasch model and the model's implicit assumption that there is no differential item functioning (DIF).

The organization of the paper is as follows. The notion and issue of DIF defined the starting point of the analysis reported in this paper. For this reason, Section 2 opens with a short section on DIF in educational testing, including quotes that show that PISA also regards DIF as a serious issue. Section 3 offers a brief overview of the data on reading in PISA 2006. Section 4 describes the scaling model used by PISA and the way PISA tests and uses the model for comparison of student scores in different countries. The rest of the paper describes our analysis of data on reading. Section 5 presents results from an analysis of data from 56 countries. In this section, focus is on the inadequacy of the Rasch model, the evidence of DIF and the effect of DIF on country comparisons. Section 6 gives results from an analysis of data from UK and Denmark, illustrating two ways to address the problems when item responses do not fit the scaling model.

2 Differential item functioning

The notion of differential item functioning (DIF) is fundamental for comparative research using summated scales. Dorans & Holland (1993) and Schmitt & Dorans (1987) describe DIF as

"… an expression which describes a serious threat to the validity of tests used to measure the aptitude of members of different populations or groups. Some test items may simply perform differently for examinees drawn from one group or another or they may measure "different things" for members of one group as opposed to members of another. Tests containing such items may have reduced validity for between-group comparisons, because their scores may be indicative of a variety of attributes other than those the test is intended to measure."

We refer to the collection of papers edited by Holland & Wainer (1993) for a comprehensive introduction to DIF.

Prais (2003) compares results on mathematics from different educational surveys and describes a situation that may have been caused by unrecognized DIF. Until the PISA survey of 2000, all surveys carried out by the International Association for the Evaluation of Educational Achievement (IEA) including a survey from 1999 "consistently pointed to low average attainment by UK pupils". The PISA survey of 2000 gave a very different result. According to PISA, the "average pupil's mathematical attainment had caught up even with Switzerland, hitherto seen as a sort of model educational system within the European context, to which England might aspire only distantly." Responding to this observation, Adams (2003) points out that IEA and PISA measure different aspects of mathematical attainment. This may or may not be true, but it cannot explain differences of this kind. Different dimensions of mathematical attainments are known to be strongly correlated, and international surveys measuring such traits should produce similar results if measurements are valid.

2.1 Confounding due to DIF

DIF can confound comparisons of countries because the outcome of the comparisons depends on the mixture of items. To see this, consider a situation where the set of items partitions into two subsets: neutral anchor items where item difficulties are the same in all countries and DIF items where difficulties depend on Country. Ordinary scaling models implicitedly assume that the ranking

of countries by the total score over anchor items is the true ranking. Notice next, that the set of DIF items partitions in different ways for different countries into items that are easier than expected by a common scaling model and items that are more difficult so that ranking of countries by the score over all items may be different from the ranking by the score over the anchor items. The degree to which confounding is important depends on several factors. If there are relatively few DIF items compared to the number of anchor items, confounding may be irrelevant, but confounding may be a problem if there are many DIF items and relatively few anchor items. Another factor is the ratio between the numbers of easy and difficult items. If there are many difficult items compared to easy items relative for a specific country, then the rank of this country will probably be lower than the true ranking by anchor items. If the large majority of DIF items are easy, then the ranking by the score over all items will probably be higher than the true ranking. A third factor is the differences in degrees of difficulties for DIF items in different countries. One item that is much easier than expected by the scaling model can outweigh several items that are a little bit more difficult than expected.

The above factors complicate assessment of the robustness of comparisons to DIF. To make matters worse, robustness also depends on the study design. In small sample studies, a limited degree of systematic bias due to DIF may be ignorable compared to the unsystematic random errors of such studies. In large sample studies with miniscule standard errors, a limited amount of DIF can have an effect on ranking. PISA is a survey with an extremely large sample n = 398,750. Collecting data for a study of this size is very expensive and it is reasonable to assume that the sample is so large in order to obtain very precise rankings also among countries where differences in educational attainments are limited. If this is true, then assessment of DIF in PISA has to be careful to a more than usual degree to make sure that confounding due to DIF really is ignorable.

2.2 Dealing with DIF

There are at least two ways to deal with DIF: by purification or by item-splitting.

Purification assumes that measurement by neutral anchor items is valid and therefore that country ranks based on scores from anchor items are estimates of true ranks. Purification attempts to identify the anchor items by a stepwise procedure where each step identifies and eliminates one or more DIF items and follows up with a test of whether there are DIF among the remaining items.

Item-splitting refers to analyses where DIF items are replaced by virtual items, one for each country, with the same contents but different item difficulties where responses to virtual items are

missing in all but one country. Ranking of countries by scores from anchor and splitted items will agree with the true ranking under two assumptions: the set of anchor items must not be empty and the scaling model must be adequate after item-splitting.

2.3 DIF in PISA

PISA also recognizes the problem of creating neutral items that function in exactly the same way in all countries.

The technical report (OECD, 2006) says that "The interpretation of a scale can be severely biased by unstable item characteristics from one country to the next" and reports (page 104) that due to DIF, "Out of 179 mathematics, reading and science items, 28 items were omitted in a total of 38 occurrences for the computation of national scores".

Kirsch et.al (2000) write that "A statistical index called *differential item functioning* was used to detect items that worked differently in some countries. … As a result, some items were excluded from scaling as if they had not been administered in that country. Table A1.4 lists the items that are excluded from the national scaling for each country".

Adams (2003), however, claims otherwise, "Item deletion occurred only where an item was found to be translated erroneously, or where there was an error in the item format or if there was a misprint in a country's booklet", and Adams et.al. (2007, page 274) say that "an item with sound characteristics in each country but that shows substantial item-by-country interactions may be regarded as a different item (for scaling purposes) in each country (or in same subsets of countries)".

All authors apparently acknowledge the problem of DIF, but do not agree on how it was dealt with by PISA. Kirsch says that DIF was taken care of by purification. Item splitting was mentioned as an option in the technical reports from 2000 and 2003, but is not mentioned in the 2006 report. Despite this, Adams et.al. (2007) claim that country DIF in PISA was dealt with by item-splitting.

3 Data on reading in PISA 2006

The reading inventory of PISA 2006 contains 28 items, which are collected in eight testlets or reading units associated with eight texts. PISA's items appear in a total of 14 booklets, but only one booklet includes all reading units, and six booklets have no reading units at all. Table 1 gives an overview of the reading units and the booklets in which they appear. The texts and items of the

reading units are not available to the public, but examples of reading units that have not been used for assessment of students can be found in OECD(2000).

Table 1. Overview of reading units in PISA 2006

Reading units	Number of items	Appears in booklets
R055 – Drugged spiders	4	6,9,11,13
R067 – Aesop	3	2,6,7,12
R102 – Shirts	3	2,6,7,12
R104 – Telephone	3	6,9,11,13
R111 – Exchange	3	6,9,11,13
R219 – Employment	3	2,6,7,12
R220 – South Pole	5	2,6,7,12
R227 – Optician	4	6,9,1,13

Note reading units R055 and R219 also appeared in a booklet referred to as Booklet 20 that was administered to 830 students from seven countries.

The eight reading units in Table 1 partition into two different sets with a total of 14 items each. Set 1 consists of R055, R104, R111 and R227 and appears in booklets 6, 9, 11 and 13. Set 2 consisting of R067, R102, R219 and R220 appears in booklets 2, 6, 7, 12. Booklet 6 is the only booklet that contains all eight reading units.

Due to the partitioning of reading units in different booklets and the fact that some booklets did not have any reading items at all, the PISA survey contains four groups of students as shown in Figure 1.

Item set 1	**Item set 2**	**Summary test results**
183,569 students Without responses to reading items		**Plausible values**
92,635 students with observed responses	**92,635 students without responses to item set 2**	**Plausible values**
91,941 students without responses to item set 1	**91,941 students with observed responses**	**Plausible Values**
30,605 students with responses to items from both item sets		**Plausible Values**

Fig. 1. Overview of data on reading in PISA 2006. Plausible values are random student scores drawn from the posterior distribution of the latent trait variable given responses to items and outcomes on conditioning variables. (See Section 4.2 for a definition of plausible values)

Table 2 contains an overview of the 28 reading items. The majority of the items are dichotomous, but a few are so-called partial credit items with integer item scores from 0 to 2. PISA distinguishes between three types of reading items: retrieving of information from texts (7 items),

interpretation (14 items), and reflection (7 items). Not all items were administered in all countries. The items are shown in Table 2. Twenty items were used in all countries and 27 items were used in both UK and Denmark.

Table 2. Reading items, PISA 2006

Item label	Item	Item type	Maximum item score	Included in all countries	Included in DK & UK	Item set 1[1]	Item set 2[1]
A	R055Q01	Interpreting	1		+	(+)	
B	R055Q02	Reflecting	1	+	+	+	
C	R055Q03	Interpreting	1	+	+	+	
D	R055Q05	Interpreting	1	+	+	+	
E	R067Q01	Interpreting	1	+	+		+
F	R067Q04	Reflecting	2	+	+		+
G	R067Q05	Reflecting	2	+	+		+
H	R102Q04A	Interpreting	1		+		(+)
I	R102Q05	Interpreting	1		+		(+)
J	R102Q07	Interpreting	1		+		(+)
K	R104Q01	Information	1	+	+	+	
L	R104Q02	Information	1	+	+	+	
M	R104Q05	Information	2	+	+	+	
N	R111Q01	Interpreting	1	+	+	+	
O	R111Q02B	Reflecting	2		+	(+)	
P	R111Q06B	Reflecting	2	+	+	+	
Q	R219Q01E	Interpreting	1		+		(+)
R	R219Q01T	Information	1		+		(+)
S	R219Q02	Reflecting	1	+	+		+
T	R220Q01	Information	1	+	+		+
U	R220Q02B	Interpreting	1				(+)
V	R220Q04	Interpreting	1	+	+		+
W	R220Q05	Interpreting	1	+	+		+
X	R220Q06	Interpreting	1	+	+		+
Y	R227Q01	Interpreting	1	+	+	+	
Z	R227Q02T	Information	2	+	+	+	
a	R227Q03	Reflecting	1	+	+	+	
b	R227Q06	Information	1	+	+	+	

[1]: (+) indicates that the item belongs to the item set but was not used in all countries.

PISA data from 2006 can be downloaded at http://pisa2006.acer.edu.au/downloads.php. The downloaded data set contains responses to reading items from 56 countries. Responses are missing for some items in some countries, but twenty items have been administered in all countries. Some students have missing responses to items and item responses are not missing at random, because the frequencies of students with at least one missing item response is very different from country to

country. In one country (Colombia), more than 25 % of the students had at least one missing item response. In other countries, 97-99 % of the students had complete responses to the twenty items. The technical report (OECD, 2006) offers no explanation for the differences, but it is obvious that items must have been administered and/or scored in different ways in different countries.

4 PISA methodology

4.1 The model

Item response theory (IRT) contains a number of statistical scaling models that describe how responses to items depend on outcomes of a latent trait variable. PISA's scaling model, often referred to as a Rasch model, is one such model.

The variables of the model are items ($Y_1,\ldots,Y_{28}$), the total score $S = \Sigma_i Y_i$, a unidimensional latent variable Θ, Country C, and other exogenous variables X.

The Rasch model makes two assumptions of conditional independence,

1) Local independence: $Y_1 \perp Y_2 \perp \ldots \perp Y_{28} \mid \Theta$
2) No DIF: $Y \perp (C,X) \mid \Theta$

and assumes that the conditional probabilities of item responses given Θ are

$$P(Y_i = y \mid \Theta = \theta) = \frac{\exp\left(y(\theta + \alpha_{iy})\right)}{\sum_{z=0}^{\max(Y_i)} \exp\left(z(\theta + \alpha_{iz})\right)} \quad (1)$$

An alternative model, which is used in other educational surveys, is the so-called 2-parameter IRT model given by

$$P(Y_i = y \mid \Theta = \theta) = \frac{\exp\left(y\beta_i(\theta + \alpha_{iy})\right)}{\sum_{z=0}^{\max(Y_i)} \exp\left(z\beta_i(\theta + \alpha_{iz})\right)} \quad (2)$$

The θ in Formulas (1) and (2) is usually referred to as the person parameter of the IRT model. The α parameters are item parameters on which the difficulties of items depend. In other words, item difficulties are independent of persons, countries and other exogenous variables. α parameters are often rewritten in terms of so-called thresholds where $\tau_{i1} = -\alpha_{i1}$ for dichotomous and polytomous items and $\tau_{i2} = \alpha_{i1} - \alpha_{i2}$ for polytomous items. The β parameter in (2) is referred to as an item discrimination parameter.

It is easily seen that S is sufficient for θ in the Rasch model and therefore that $Y \perp \Theta \mid S$. The same is not true for the 2-parameter model. There are many advantages of having a sufficient score in terms of simplified statistical inference and a wider range of ways to check the adequacy of the model, but the choice of this model is not justified without evidence supporting the claim that the item discrimination is the same for all items.

The PISA model assumes that Θ has a conditional normal distribution with means depending on C and X. Under this model, comparison of countries appears to be a straightforward exercise in latent structure analysis.

4.2 Assessing the fit of item responses to Rasch models

According to OECD (2006), "particular attention was paid to the fit of the items to the scaling model" during PISA's analysis of data. This was done in three ways: by calculation of so-called Infit coefficients; by calculation of Discrimination coefficients; and by comparison of estimates of item parameters in different countries with estimates of item parameters in a subsample of 15,000 students from 30 OECD countries. The analysis of fit of items to PISA's scaling model is briefly described and illustrated on pages 147-152 of the technical report (OECD, 2006). The results were included in a number of tables in national reports. From the examples of these tables, it can be deduced that Infits and Discrimination coefficients have also been calculated for the subsample of 15,000 students that provided the item parameter estimates used for the international comparisons. These results, which could have served as evidence supporting the adequacy of PISA's scaling model, are not found in the technical report.

4.2.1 Infits

Let Y_{vi} be the observed item score for person v on item Rasch and let E_{vi} be the expected score. The Infit measure is sometimes referred to as a "weighted" sum of squared residuals, but is in fact just the sum of squared *unstandardized* residuals divided by the sum of the variances of the residuals,

$$Infit_i = \frac{\sum_v (Y_{vi} - E_{vi})^2}{\sum_v Var(Y_{vi} - E_{vi})} \quad (3)$$

The expected item score and the variance of the residuals can easily be derived from (1). In practice, item parameters are unknown and therefore replaced by estimates of the parameters. The

expected Infit is claimed to be 1. In PISA, there is no formal assessment of the significance of the Infits. They are, however, reported in such a way that it appears as if values outside the interval [0.7,1.3] are regarded as evidence of inadequate fit of the item to the Rasch model. If the person parameter estimates had been consistent, it would probably have been correct that the expected Infit is asymptotically equal to 1. The person parameter is, however, not consistent unless it can be assumed that the number of items goes towards infinity. Kreiner & Christensen (2011a) show that residuals and functions of residuals are systematically biased when the number of items is constant and the person parameter estimate is plugged into the formula for the expected item scores. They also show that residuals and Infits are unbiased if the probabilities in Formula (1) are replaced by the conditional probabilities of item responses given the total score on all items, and show how to assess the significance of the departure of the conditional Infits from 1.

4.2.2 Discrimination coefficients

PISA uses the point-biserial correlation coefficient between an item and the total score on all items as a measure of discrimination and requires the point-biserial discrimination to be higher than 0.25. They do not explain where this threshold originated and why a lower coefficient should be regarded as evidence against the scaling model. Their analysis of discrimination coefficients also misses the important point that discrimination coefficients should be used to check the Rasch model's claim that all items discriminate in the same way.

4.3 Country comparisons

PISA uses a four-step procedure for their analyses of country differences.

*Step 1. National item calibration***.** The parameters of the Rasch model and the parameters of the conditional distribution of Θ are estimated separately for each country.

*Step 2. International item calibration***.** The item parameters of the Rasch model are estimated on a random subset of 15,000 students from 30 OECD countries. These estimates are used as estimates of item parameters that apply for all countries during the generation of plausible values in the next step. Exogenous variables are not taken into account during this step. It is not clearly stated, but we assume that the analysis here is made under the assumption that Θ has a marginal normal distribution in each country. There are, in other words, subtle differences between the scaling models of Step 1 and the scaling model of Step 2.

Step 3. Calculations of plausible values. Plausible values are random numbers drawn from the posterior distribution of Θ given (Y,X). We refer to the variable generated in this way as $\hat{\Theta}$. The posterior distribution is defined by the estimates of the item parameters obtained during the international item calibration. The parameters of the conditional distribution of Θ given X come from the national item calibration. It is not clear whether these parameters have been re-estimated under the scaling model based on the common item parameter estimates from Step 2.

Step 4. Analysis of variance. PISA finally compares countries by a one-sided ANOVA of $\hat{\Theta}$ given C.

Using plausible values may appear to be an attractive way to do a latent structure analysis based on Rasch models for two reasons. The first is that plausible values can be drawn from the conditional distribution of $\Theta \mid S,C,X$ instead of the more complicated distribution of $\Theta \mid Y,X$. The second is that the conditional distribution of $\hat{\Theta} \mid X$ (that is $\hat{\Theta} \mid S,X$ marginalized over S) is the same as the distribution of $\Theta \mid X$

$$\Theta \mid X \sim \text{Norm}(\xi,\sigma^2) \Rightarrow \hat{\Theta} \mid X \sim \text{Norm}(\xi,\sigma^2) \quad (4)$$

5 ANALYSIS 1: Data on reading in 56 countries

The analysis reported in this section used data from 28,593 students with complete responses to the twenty reading items that were administered in all countries. All students were exposed to Booklet 6, since this was the only booklet that contained all reading items. Appendix A offers country information on the number of students and the average scores on the twenty items.

The analysis in this section attempts to replicate the analysis carried out by PISA. We calculate Infit and discrimination coefficients and compare item parameters estimated in different countries to item parameters estimated in the complete data set for an analysis of DIF. There are, however, also differences. These are described below.

The type of inference in PISA where item and population parameters are jointly estimated is often referred to as marginal inference. Our analysis is conditional focusing on the conditional distribution of items given the total score on all items. Conditional inference does not depend on assumptions on the distribution of Θ and is therefore, in our opinion, generally preferable to marginal inference. It is known, however, that marginal estimates of item parameter estimates are

relatively robust in situations where the distributional assumption underlying marginal inference is incorrect. For this reason, we do not expect the differences in the inference paradigms between PISA's and our analyses to be important.

A second difference between PISA's and our analyses is that we use Andersen's (1973) conditional likelihood ratio (CLR) to test the hypothesis that item parameters are the same for students with low scores and students with high scores.

A third difference is that the significance of the departure between observed fit coefficients and the expected coefficients under the Rasch model is assessed for all coefficients. We always adjust for multiple testing during such analyses, but it turned out to be of no consequence in this case because all p-values are very close to zero. This will not be further addressed here.

Fourth, the Infit used during our analysis differs from the Infit used by PISA. The Infit coefficient that we use compare observed item scores to expected item scores score in the conditional distribution of items given the student's score on all items. The advantages of this procedure are that it avoids the bias inherent in the traditional Infits and that it is easy to derive the asymptotic distribution of the Infit from the central limit theorem. See Kreiner & Christensen (2011a) for details on conditional inference of residuals in Rasch models.

Fifth, we use a different discrimination coefficient than the one used by PISA. Instead of the point-biserial correlation, we use Goodman & Kruskal's (1954) γ to measure the degree of association between item scores and restscores on the other items because it is more appropriate for analysis of association among ordinal categorical variables. Another and more important difference between PISA's and our analyses is that we calculate the expected coefficient under the Rasch model and assess the significance of the difference between the observed and expected coefficients (see Kreiner (2011) for additional information on analysis of item-restscore association).

Finally, our analysis of DIF uses stronger and more appropriate fit statistics than PISA. An overall analysis of DIF is provided by Andersen's CLR test of the hypotheses that item parameters are the same in all countries against the alternative that each country has its own set of item parameters. In addition to this, we calculated two different tests of no DIF for separate items. The first of these is a χ^2 test of conditional independence of separate items and country given the total score on all items. Such hypotheses are often used during DIF analyses for items from all kinds of IRT models, but the Rasch model is the only model where the hypothesis of conditional independence is actually true because it is the only model with a sufficient score. The second test is

a CLR test of the Rasch model against an alternative where item parameters for one specific item are different in different countries (Kelderman, 1984; 1989).

5.1 Results

Figure 2 plots the conditional maximum likelihood estimates of item parameters based on the sample of 28,953 students with complete responses to the 20 items against the marginal estimates reported by PISA (OECD, 2006; Appendix A) that were based on a random sample of 15,000 students from OECD countries. There are a few items with minor discrepancies, but the general impression is that estimates are consistent.

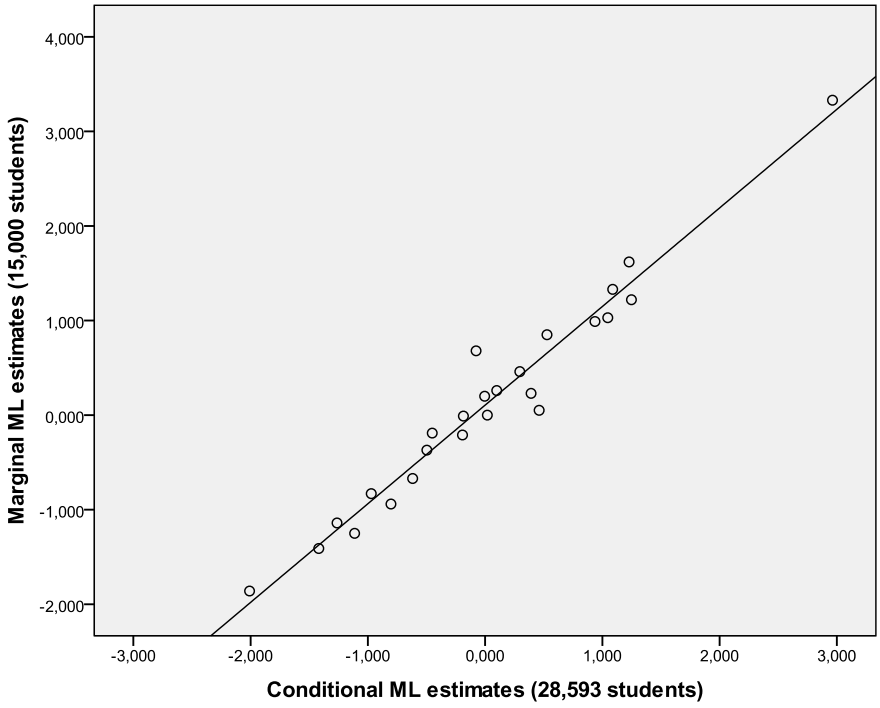


Fig. 2. Association between conditional maximum likelihood estimates and marginal estimates of item parameters (thresholds) of the twenty items that were administered in all countries.

Figure 3, plots the ranks of countries given by the total scores over the twenty items against the ranks published by PISA (OECD, 2007). In this plot, the consistency between PISA's results and the results obtained by analysis of Booklet 6 data is less convincing. It is possible that the differences in rankings to some extent can be explained if the subsample of students exposed to Booklet 6 is not representative and by the fact that our ranking is based on students with complete responses to all twenty items. The fact that the Booklet 6 ranking is based on scores over twenty

items whereas PISA's ranking include all items that were administered in the countries can also explain some of the differences in rankings. Apart from this, the Rasch model expects that the rankings by the different subsets of items should be similar.

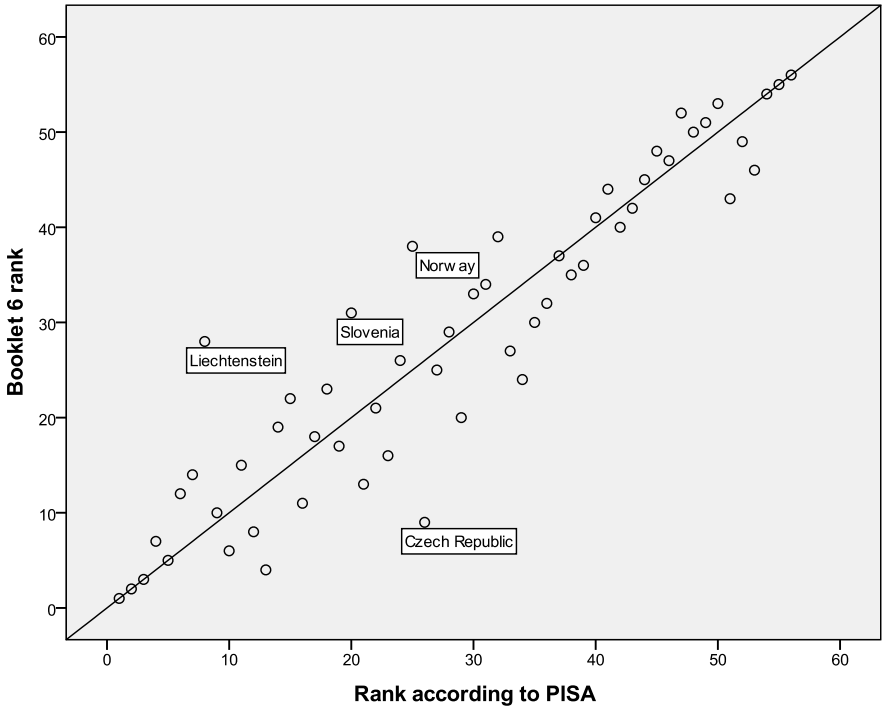


Fig. 3. Country ranks defined by the total score over twenty items for students that responded to items in Booklet 6 against the ranks of countries according to PISA

Figure 3 therefore suggests that something is not as it should be. That this is so is confirmed by our analysis of the fit of the Rasch model to the responses of the twenty items. The evidence against the Rasch model is overwhelming. The CLR test rejects the hypothesis that item parameters are the same among students with scores from 1-13 and students with scores from 14-24 (CLR = 5371,0; df = 24; $p < 0.00005$) and the hypothesis that item parameters are the same in all countries (CLR = 27,389.0; df = 1320, $p < 0.00005$). Table 3, which shows the item fit statistics for separate items, tells the same story. The Infit accepts three items, two of which are also accepted by the γ coefficients measuring item-restscore association, but all items are rejected by the tests for DIF. Taken as a whole, only conclusion can be drawn from this analysis: nothing here remotely resembles a Rasch model.

In item response theory, item characteristic curves (ICC) are functions describing the expected score on items as functions of the person parameter θ of an IRT model. The estimated ICC of item R055Q03 under the Rasch model is shown in Figure 4 together with average items scores at

the different values of the person parameter estimates for all students (black) and for student from UK (green) and Denmark (red). In this figure, the black points are points on the empirical ICC. This curve is steeper than the ICC under the Rasch model, suggesting that the item discrimination of this item is stronger that expected by the Rasch model. This agrees with the results on this item in Table 3 where the observed item-restscore association ($\gamma = 0.633$) is significantly stronger than the expected association ($\gamma = 0.537$). DIF is illustrated by empirical ICC curves for UK (green) and Denmark (red). R055Q03 appears to be systematically easier in UK and systematically more difficult in Denmark compared to the population of students as a whole.

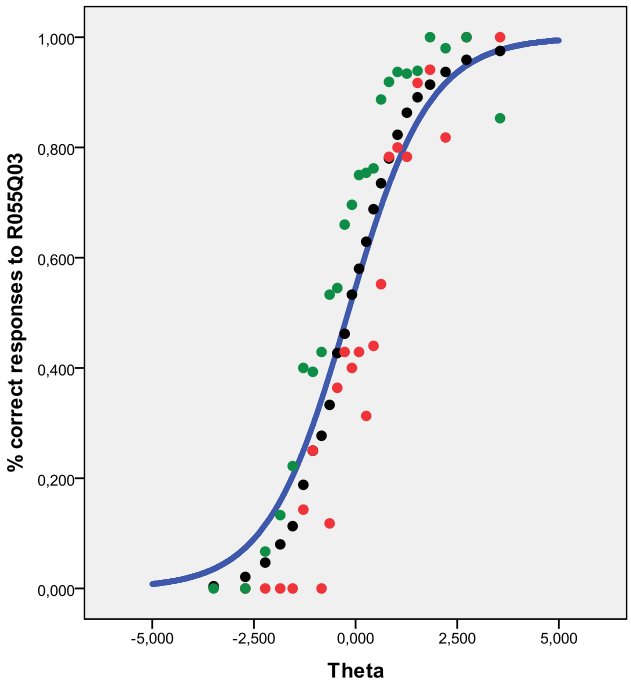


Fig. 4. Estimated ICC curve for R055Q03 together with points on the empirical ICC curves for all students (black), UK students (green) and Danish students (red)

During the calculation of the CLR test of no DIF, observed and expected item scores in are compared on a country wise basis. Information on items that are significantly easier or more difficult in different countries can be found in Appendix B. All items appear in the lists of easy and difficult items for several countries, so that the set of anchor items without DIF appears to be empty.

Table 3. Estimates of item parameters and item fit statistics based on Booklet 6 data on 20 PISA items.

				Item-restscore gamma								
item	thresholds		infit	p	observed	expected	p	chi	df	p	clr	df p
R055Q02	0.461	.....	0.968	0.000	0.554	0.522	0.000	896.1	684	0.000	680.2	55 0.000
R055Q03	-0.185	.....	0.894	0.000	0.633	0.537	0.000	1099.3	684	0.000	1213.2	55 0.000
R055Q05	-0.971	.....	0.822	0.000	0.719	0.559	0.000	1230.9	684	0.000	630.5	55 0.000
R067Q01	-2.009	.....	0.978	0.055	0.601	0.590	0.154	1208.6	684	0.000	891.3	55 0.000
R067Q04	-0.451	0.528	1.242	0.000	0.457	0.578	0.000	2644.4	1316	0.000	3230.5	110 0.000
R067Q05	0.296	-0.777	1.233	0.000	0.544	0.643	0.000	2913.7	1316	0.000	3515.3	110 0.000
R104Q01	-1.419	.....	0.899	0.000	0.681	0.572	0.000	1373.6	684	0.000	885.8	55 0.000
R104Q02	1.088	.....	1.166	0.000	0.325	0.512	0.000	1335.4	684	0.000	1284.7	55 0.000
R104Q05	0.937	2.963	0.998	0.782	0.563	0.531	0.000	2997.6	1103	0.000	2222.3	110 0.000
R111Q01	-0.497	.....	0.971	0.000	0.586	0.545	0.000	1049.4	684	0.000	594.9	55 0.000
R111Q06B	1.228	0.019	1.123	0.000	0.565	0.620	0.000	1883.8	1262	0.000	2231.8	110 0.000
R219Q02	-1.263	.....	0.891	0.000	0.663	0.567	0.000	1488.2	684	0.000	1104.4	55 0.000
R220Q01	1.046	.....	0.871	0.000	0.653	0.513	0.000	1108.5	632	0.002	868.7	55 0.000
R220Q04	-0.004	.....	1.003	0.545	0.539	0.532	0.253	932.5	684	0.000	717.3	55 0.000
R220Q05	-1.113	.....	0.923	0.000	0.670	0.563	0.000	994.8	684	0.000	368.7	55 0.000
R220Q06	-0.193	.....	1.055	0.000	0.498	0.537	0.000	1044.4	684	0.000	1075.3	55 0.000
R227Q01	0.392	.....	1.132	0.000	0.398	0.524	0.000	1183.3	684	0.000	1494.5	55 0.000
R227Q02T	-0.803	1.248	1.099	0.000	0.501	0.559	0.000	2856.8	1316	0.000	3359.9	110 0.000
R227Q03	0.099	.....	0.913	0.000	0.615	0.530	0.000	1216.9	684	0.000	656.5	55 0.000
R227Q06	-0.618	.....	0.862	0.000	0.679	0.548	0.000	1354.7	684	0.000	1500.0	55 0.000

5.2 Consequences of DIF

Inference based on an erroneous model can be robust to model errors, in which case we would consider model errors as irrelevant. To address the relevance of the lack of fit between PISA's data and PISA's scaling model, we examine the degree to which country rankings depend on the choice of items. Recall that DIF among items can have the effect that the rankings of countries by scores over different subsets of items differ but that it does not have to be so. If it turns out that, except for unsystematic random errors, country rankings are the same for all subsets of items; one would be justified in arguing that the DIF among items is irrelevant. If rankings are very different, DIF is indisputably relevant.

The first part of Table 4 shows the ranking according to the total score on the twenty items. According to this ranking, UK is no. 23 and Denmark no. 17.

The second part of Table 4 shows the ranking of countries according to scores from the following four subsets of items where the observed items scores in UK and Denmark departed significantly from the item scores expected by the Rasch model (see Appendix B for similar information on other countries):

Subset 1: R055Q02, R055Q03, R104Q01, R104Q02, R219Q02, R227Q06

Subset 2: R067Q04, R111Q06B, R220Q01, R220Q04, R220Q05, R220Q06

Subset 3: R055Q02, R104Q01, R111Q01, R219Q02, R220Q05, R220Q06, R227Q01, R227Q02T, R227Q06

Subset 4: R055Q03, R067Q04, R067Q05, R104Q05, R220Q04, R227Q03

For UK and Denmark, the differences in rankings are dramatic. Subset 1 puts UK close to the top of the countries, while Subset 3 relegates UK to no. 36 out of the 56 countries. Ranking by Subset 3 posits Denmark as no. 3 right after Korea and Finland with UK as no. 24, but Denmark drops to no. 42 if Subset 4 is used to rank countries.

Ranking errors depend on the reliability of the scales. It is therefore no surprise that ranking by subsets of items differs from rankings by the complete set of items since ranking errors are more probable for the subsets with relatively few items. Appendix C describes different ways to assess the ranking errors according to total scores and subscores under the Rasch model. We have used these methods to evaluate the degree to which the rankings in Table 4 contradict what the Rasch model expects. The results are as follows.

Using all twenty items, the probability that the rank of UK is in the [19,24] interval is 0.949. For Denmark, the probability that the rank is in [14-20] is equal to 0.943. Under the Rasch model, the probability that the Subset 1 rank of UK is equal to 8 or higher is equal to 0.0082, and the probability that the Subset 2 rank is 36 or lower is 0.0086. The probability that Denmark is ranked among the first three countries by Subset 3 is 0.0030, while the probability that Denmark is number 42 or lower if Subset 4 is used for ranking is equal to 0.00001. The results confirm that the rankings by the subsets are beyond what we would expect from the Rasch model and, therefore, that it cannot be ignored that the effect of the DIF in PISA's items of the ranking of countries could confound the ranking of countries.

Notice also that what we see in Table 4 is exactly what Prais (2003) saw in IEA's and PISA's assessment of mathematics in 1999 and 2000: that different sets of items may result in dramatically different rankings of countries. Whether DIF lies behind the findings of Prais is, of course, a question yet unanswered, but the similarities between the situations are striking.

Another way to assess the effect of DIF is to compare response frequencies on items in different score groups where the Rasch model expects the relative frequencies to be the same in all countries. Table 5 shows the frequencies for two partial credit items where responses among Danish students are extreme compared to other countries. It just so happens that responses to the same items are just as extreme among students from Colombia (although in a different way), for which reason we include response frequencies for Colombia together with response frequencies from UK and Denmark. To limit the size of the table, we only include response frequencies from three score groups, but we would like to stress that what can be seen in Table 5 are systematic differences appearing (more or less pronouncedly) in all score groups.

Table 4. Ranking of countries by different subsets of items

	Twenty items		Subset 1		Subset 2		Subset 3		Subset 4	
Rank	Country	Mean	Country	Mean	Country	Mean	Country	Mean	Country	Mean
1	Korea	17.52	Finland	4.41	Finland	5.40	Korea	7.79	Korea	5.95
2	Finland	17.12	N. Zealand	4.29	Korea	5.37	Finland	7.68	Hong Kong	5.58
3	Hong Kong	16.20	Korea	4.27	Estonia	4.89	**Denmark**	**7.65**	Finland	5.56
4	Estonia	15.79	France	4.21	Hong Kong	4.83	Netherland	7.37	Taipei	5.51
5	N. Zealand	15.35	Netherland	4.16	Canada	4.65	Hong Kong	7.28	Macao-Chin	5.16
6	Netherland	15.33	Canada	4.14	Macao-China	4.58	Sweden	7.09	Canada	5.01
7	Canada	15.20	Hong Kong	4.10	Czech rep.	4.54	Germany	7.08	Poland	4.87
8	Belgium	15.09	**UK**	**4.09**	Latvia	4.54	Switzerland	6.96	Estonia	4.80
9	Czech rep.	15.03	Belgium	4.06	N. Zealand	4.53	Austria	6.94	N. Zealand	4.80
10	Poland	15.01	Ireland	4.06	Belgium	4.50	N. Zealand	6.93	Portugal	4.79
11	Taipei	14.94	Australia	4.05	Ireland	4.50	Belgium	6.91	Latvia	4.76
12	Ireland	14.94	Czech rep.	4.03	Netherland	4.47	Japan	6.90	Ireland	4.72
13	Macao-Chin	14.80	Estonia	4.03	Taipei	4.45	Lichtenstein	6.81	Belgium	4.69
14	Australia	14.76	**Denmark**	**3.99**	Poland	4.45	Ireland	6.77	Czech rep.	4.68
15	Sweden	14.67	Poland	3.93	France	4.41	Australia	6.76	Australia	4.67
16	France	14.6	Taipei	3.90	Germany	4.37	Czech rep.	6.76	Greece	4.67
17	**Denmark**	**14.59**	Sweden	3.90	Hungary	4.37	Canada	6.70	Chile	4.64
18	Germany	14.57	Spain	3.85	Japan	4.37	Estonia	6.68	Netherland	4.63
19	Switzerland	14.46	Switzerland	3.84	Slovak rep.	4.32	France	6.67	France	4.60
20	Latvia	14.45	Portugal	3.81	Italy	4.29	Slovenia	6.61	**UK**	**4.60**
21	Austria	14.32	Japan	3.8	Sweden	4.27	Poland	6.56	Turkey	4.58
22	Japan	14.27	Austria	3.78	Greece	4.25	Luxembourg	6.49	Slovak rep.	4.57
23	**UK**	**14.23**	Latvia	3.70	Australia	4.22	Macao-Chin	6.47	Italy	4.54
24	Slovak rep.	13.96	Slovenia	3.70	Chile	4.20	**UK**	**6.45**	Germany	4.50
25	Hungary	13.94	Iceland	3.69	Portugal	4.18	Hungary	6.41	Russian fed.	4.48
26	Iceland	13.91	Germany	3.67	Croatia	4.17	Iceland	6.41	Hungary	4.43
27	Italy	13.83	Macao-Ci	3.65	Switzerland	4.16	Taipei	6.38	Switzerland	4.41
28	Lichtenstein	13.78	Lichtenstein	3.63	**Denmark**	**4.15**	Norway	6.36	Lituania	4.34
29	Luxembourg	13.69	Italy	3.58	Spain	4.15	Latvia	6.29	Spain	4.34
30	Spain	13.66	Slovak rep.	3.54	Austria	4.12	Spain	6.24	Luxembourg	4.30
31	Slovenia	13.61	Chile	3.5	Luxembourg	4.09	Greece	6.16	Sweden	4.28
32	Greece	13.6	Luxembourg	3.5	Iceland	4.05	Italy	6.12	Austria	4.27
33	Croatia	13.49	Hungary	3.49	Lichtenstein	4.00	Croatia	6.06	Croatia	4.27
34	Portugal	13.48	Greece	3.42	Lituania	3.99	Slovak rep.	6.05	Iceland	4.20
35	Chile	13.10	Russian fed.	3.32	Slovenia	3.93	Russian fed.	5.88	Japan	4.04
36	Russian fed.	13.08	Lituania	3.31	**UK**	**3.91**	Portugal	5.73	Colombia	4.01
37	Turkey	12.96	Uruguay	3.26	Norway	3.89	Chile	5.68	Uruguay	3.99
38	Norway	12.91	Norway	3.25	Turkey	3.87	Lituania	5.48	Israel	3.93
39	Lituania	12.88	Israel	3.24	Uruguay	3.85	Israel	5.41	Slovenia	3.86
40	Uruguay	12.38	Croatia	3.19	Russian fed.	3.63	Uruguay	5.33	Lichtenstein	3.81
41	Israel	12.17	Turkey	3.18	Israel	3.62	Turkey	5.32	Mexico	3.80
42	Mexico	11.33	Mexico	3.12	Mexico	3.56	Mexico	4.91	**Denmark**	**3.66**
43	Colombia	10.88	Bulgaria	2.73	Colombia	3.55	Thailand	4.87	Thailand	3.59
44	Thailand	10.88	Colombia	2.71	Argentina	3.25	Bulgaria	4.65	Brazil	3.49
45	Bulgaria	10.45	Argentina	2.70	Jordan	3.25	Serbia	4.53	Bulgaria	3.45
46	Argentina	10.17	Serbia	2.53	Thailand	3.19	Jordan	4.45	Argentina	3.44
47	Serbia	9.90	Thailand	2.53	Brazil	2.99	Argentina	4.29	Norway	3.43
48	Jordan	9.82	Romania	2.33	Tunisia	2.94	Indonesia	4.27	Serbia	3.18
49	Tunisia	9.33	Brazil	2.32	Bulgaria	2.93	Colombia	4.25	Tunisia	3.15
50	Brazil	9.18	Jordan	2.29	Indonesia	2.78	Tunisia	4.07	Jordan	3.14
51	Indonesia	8.72	Indonesia	2.23	Serbia	2.71	Montenegro	3.95	Romania	3.00
52	Romania	8.54	Montenegro	2.22	Romania	2.53	Brazil	3.69	Montenegro	2.73
53	Montenegro	8.44	Tunisia	2.13	Montenegro	2.24	Romania	3.49	Indonesia	2.57
54	Azerbaijan	6.85	Azerbaijan	2.10	Azerbaijan	2.03	Azerbaijan	2.83	Azerbaijan	2.18
55	Qatar	5.39	Qatar	1.23	Qatar	1.60	Qatar	2.66	Qatar	1.52
56	Kyrgyzstan	4.87	Kyrgyzstan	1.09	Kyrgyzstan	1.56	Kyrgyzstan	2.27	Kyrgyzstan	1.51

Table 5. DIF effects. Relative distribution across item scores 0, 1 and 2 on R067Q04 and R227Q02T in different score groups in DK, UK and Colombia.

			R067Q04			*R227Q02T*		
Score	Country	n	0	1	2	0	1	2
1-5	DK	12	.91	.08	.00	.58	.33	.08
	UK	83	.77	.14	.09	.76	.22	.02
	Colombia	37	.78	.14	.08	.95	.05	.00
	All countries	3741	.81	.15	.03	.71	.26	.02
6-10	DK	68	.58	.32	.09	.15	.53	.32
	UK	180	.50	.41	.09	.44	.41	.14
	Colombia	85	.27	.45	.28	.86	.14	.00
	All countries	5881	.45	.38	.17	.46	.46	.08
11-15	DK	108	.49	.42	.09	.07	.38	.55
	UK	267	.26	.49	.24	.22	.59	.18
	Colombia	73	.20	.33	.47	.56	.42	.01
	All countries	7456	.26	.43	.31	.26	.56	.18
16-20	DK	134	.22	.48	.29	.01	.24	.75
	UK	244	.09	.44	.46	.08	.51	.42
	Colombia	41	.05	.42	.54	.29	.63	.07
	All countries	8381	.12	.38	.50	.09	.53	.38
21-24	DK	35	.03	.26	.71	.00	.14	.86
	UK	135	.02	.25	.73	.02	.27	.72
	Colombia	9	.00	.11	.89	.22	.33	.44
	All countries	2884	.02	.20	.77	.01	.33	.66

6 ANALYSIS 2: Comparison of Denmark and UK

The result of the analysis of data from all 56 countries rejected a common Rasch model for all countries. In this section, we pursue the less ambitious goal to attempt to find an adequate scaling model for students from UK and Denmark. We do this for several reasons.

The first is that the evidence that all items function differentially relative to the 56 countries does not automatically mean that items function differently in all countries. We do expect DIF for some items relative to UK/Denmark, but we also expect that some items function in the same way in both countries since UK and Denmark are relatively similar in terms of culture, language and alphabetic systems.

The second reason is that we also want to address the problem of local dependence among items. We suspect that some of the evidence of differential item discrimination is caused by local dependence, and we expect in particular that items from the same reading units could be positively locally dependent. The amount of DIF in the analysis of data from all countries frustrated serious attempts to untangle local dependence, because the DIF confounds associations among items. In a

less ambitious analysis covering two countries like UK and Denmark, there is no reason why a careful analysis of local dependence should not be feasible.

Finally, if DIF relative to UK/Denmark is limited and if some evidence of local dependence emerges, there are two ways to proceed. One of these ways is to model departures from the Rasch model so that DIF and local dependence can be adjusted for when countries are compared. The other is to purify the set of items by eliminating those items that disagree with the Rasch model.

6.1 Analysis by Rasch models

Except for the fact that the analysis included all 27 items that were administered in UK and Denmark, the initial analysis was carried out in the same way as the analysis of data from all countries and lead to exactly the same results. The conditional likelihood ratio tests comparing item parameters among students with low and high scores rejected the model ($CLR = 276.3$; $df = 32$; $p < .0005$) as did the test comparing item parameters in UK and Denmark ($CLR = 678.0$; $df = 32$; $p < .0005$). Item-restscore coefficients departing significantly from the expectation of the Rasch model and /or evidence of DIF were found for 18 items (results not shown).

In addition to these results, Kelderman's (1984) conditional likelihood ratio test of local dependence rejected local independence for 109 out of 351 pairs of items.

6.2 Modelling DIF and local dependence by graphical loglinear Rasch models

Graphical loglinear Rasch models (GLLRM) are generalizations of Rasch models where the DIF and local dependence are permitted as long as they are uniform. GLLRMs are chain graph models (Lauritzen, 1996) in which a scaling model containing a latent trait variable and a set of items has been embedded. Contrary to ordinary chain graph models, GLLRMs are characterized not by one but by two independence or Markov graphs; one without the total score called the IRT graph and one with the total score referred to as a Rasch graph. The scaling model that describes how items depend on the latent trait and other variables is similar to an ordinary Rasch model, except that loglinear interaction parameters relating to items and exogenous variables have been added to the joint conditional distribution of item scores given the latent trait variable. The assumption that DIF and LD are uniform means that the interaction parameters do not depend on the outcome of the latent variable. The Rasch graph in which the total score has been included has global Markov properties that can be very useful during the analysis of the adequacy of the model. We refer to

Kelderman (1984 and 1989) and Kreiner & Christensen (2002, 2004, 2006 and 2011b) for additional information on these models.

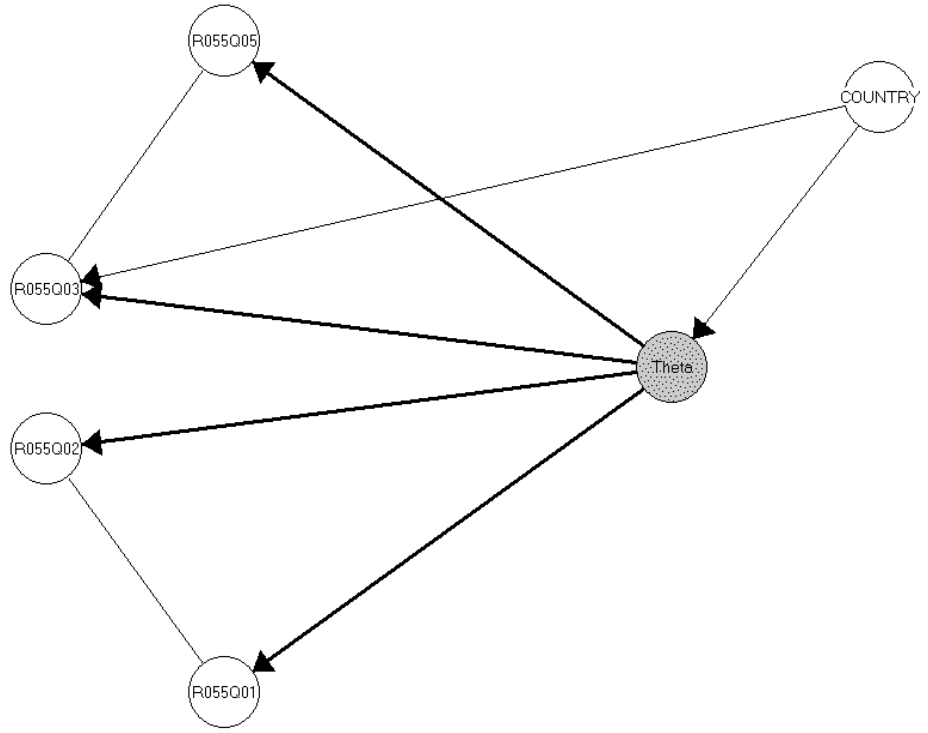
Formula (5) and Figure 5 define a GLLRM for Reading unit R055 ("Drugged spiders"). This unit has four items, R055Q01 (A), R055Q02 (B), R055Q03 (C), and R055Q05 (D). The analysis of these items by GLLRMs found no evidence against a model with two pairs of positively dependent items (R055Q01 & R055Q02 and R055Q03 & R055Q05) and one DIF item (R055Q03).

The conditional probability of responses to these items given the latent trait and country is

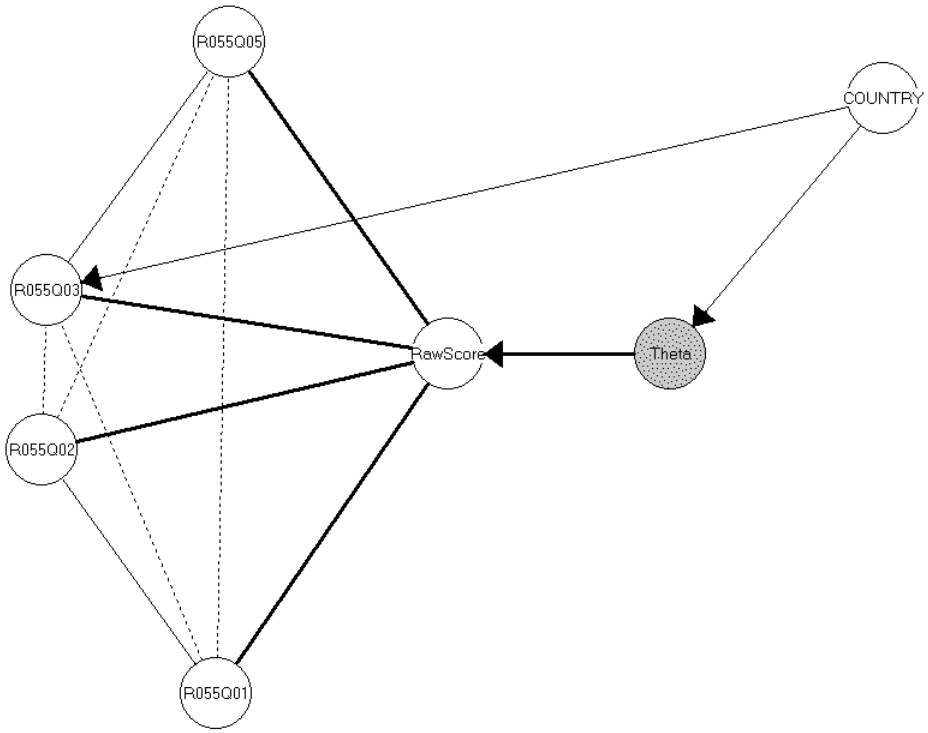
$$P(A=a,B=b,C=c,D=d \mid \Theta=\theta, Country=x) = \frac{\exp\left(s\theta-\alpha_a^{(A)}-\alpha_b^{(B)}-\alpha_c^{(C)}-\alpha_d^{(D)}+\lambda_{ab}^{(AB)}+\lambda_{cd}^{(CD)}+\delta_{cx}^{(C)}\right)}{\sum_{(i,j,k,l)}\left(\exp\left(t\theta-\alpha_i^{(A)}-\alpha_j^{(B)}-\alpha_k^{(C)}-\alpha_l^{(D)}+\lambda_{ij}^{(AB)}+\lambda_{kl}^{(CD)}+\delta_{kx}^{(C)}\right)\right)} \quad (5)$$

where s = a+b+c+d and t = i+j+k+l. Under this model, (A,B) and (C,D) are locally dependent pairs of items and C is a DIF item relative to Country, but local dependence and DIF are uniform because the interaction parameters describing these associations do not depend on θ.

The advantages of the graphical Rasch and loglinear Rasch models are that exogenous variables like Country are included in the model so that a unified treatment of different kinds of departures from the Rasch models is easier. Under the model shown in Figure 3 where R055Q03 and R055Q05 are locally dependent and where R055Q03 functions differently in UK and Denmark, a less than careful analysis of the adequacy of the Rasch model would suggest that R055Q05 also functioned differently. With country included in the model in a formal way, it is much easier to disentangle evidence of DIF from evidence of local dependence. In addition, because the model is graphical, it is possible to derive hypotheses of local independence among manifest variables of the model that are true if the model is adequate. These hypotheses follow from the global Markov properties (GMP) of the Rasch graph. One such GMP hypothesis is that R055Q05 ⊥ Country | S, R055Q03. Tests of such hypotheses and similar ones addressing assumptions of local independence provide additional powerful ways to examine whether items fit the GLLRM.



IRT graph



Rasch graph

Fig. 5 The Markov graphs of the GLLRM for R055 items. The dotted edges in the Rasch graph refer to conditional dependence that is induced by conditioning with the total score

In GLLRMs, the total score is sufficient, and conditional inference can be carried out in exactly the same way as for ordinary Rasch models. The Infits and Discrimination coefficients under this model (Table 6) found no faults, and the conditional likelihood ratio test comparing item parameters for students with low scores and item parameters for students with high scores accepted the model (CLR = 4,38; df = 6; p = 0.62), as did the test comparing item parameters in UK and Denmark (CLR = 1.74; df = 4; p = 0.78).

Table 6. Item fit statistics for under the GLLRM for R055
based on Booklet 6 data from UK and Denmark

			Item-restscore gamma					
item	infit	p	observed	expected	p	clr	df	p
R055Q01	1.028	0.71	0.614	0.652	0.35	0.98	1	0.32
R055Q02	0.978	0.51	0.573	0.553	0.55	0.04	1	0.85
R055Q03	1.025	0.55	0.631	0.645	0.65	58.54	1	<0.00005
R055Q05	0.988	0.82	0.726	0.714	0.66	1.22	1	0.27

Table 6 contains strong evidence of DIF of R055Q03 relative to country. Hypotheses of local independence are also tested by conditional likelihood ratio tests. The evidence of local dependence provided by these tests is weaker than the evidence of DIF, but nevertheless significant (R055Q01 & R055Q02: CLR = 4,48; df = 1; p = 0.0342. R055Q03 & R055Q05: CLR = 5.87; df = 1; p = 0.0154). Analysis of the fit of a GLLRM without local dependence rejected the model and resulted in significant Infits and discrimination coefficients. For this reason, we conclude that there is local dependence among some of the items of R055.

Similar analyses with similar results were carried out for the other reading units. The results are summarized in Table 7. Local dependence was found in five out of eight reading units. In two cases, the overall test of no DIF above and beyond the DIF modelled by the interaction parameters of the model rejected the model. Analysis of separate items did not indicate specific problems, and Infits and discrimination coefficients did not disagree with the GLLRM in these cases. For this reason, we accept the GLLRM in these two cases also, even though we recognize that there is room for improvement of the scaling model. The DIF in these models favoured UK in R055, R067, R111, and R219. In R102, R104, and R220, the DIF items were relatively easier in Denmark than in UK. Finally, in R227, one item (R227Q03) was easier in UK than in Denmark, while another item (R227Q06) was easier in Denmark than in UK.

The fit of GLLRMs to the separate reading units was encouraging, but that does not automatically imply that the GLLRM combining the separate reading unit models into one model can be expected to fit the data, since such a model assumes that items from different reading units are locally independent and that there is no differential reading unit functioning. For this reason, the next step was to attempt to fit a GLLRM model to the complete set of items, using the procedure for item screening described by Kreiner & Christensen (2011b) as a starting point, followed by stepwise elimination and addition of interaction terms in order to improve the fit of the model.

The search for an adequate GLLRM was abandoned at the point where the model shown in Figure 6 was reached. The model appeared to capture DIF for 15 items and describes a fairly complicated dependence structure, but a check of this model showed that it nevertheless was inadequate with very strong evidence of negative local dependence between items from different reading units. Adding local dependence to the model led to computational problems without producing a model with an adequate fit to data. For this reason, the search for a GLLRM adequate for all items was abandoned.

The failure to fit a GLLRM to all items does not rule out the possibility of a core of items that does fit such a model or even a pure Rasch model. To pursue this idea, we turn to purification in the next subsection.

6.3 Item purification by exploratory item screening

During purification, attempts are made to identify and eliminate items that do not fit the scaling model. The natural starting point for such an exercise should always be a renewed attempt to analyze the contents and formats of items in order to find errors that have been overlooked. Since the content of PISA items is a well-kept secret, we have to approach it differently.

One way to do this is to use the following modification of Kreiner & Christensen's (2011b) item screening procedure.

Table 7. Texts with reading items for PISA 2006

Reading unit	Locally dependent items	DIF items	CLR test comparing item parameters in score groups	CLR test comparing item parameters in countries
R055	R055Q01 & R055Q02 R055Q03 & R055Q05	R055Q03	CLR = 4.4; df = 6; p = 0.62	CLR = 1.7; df = 4; p = 0.78
R067	-	R067Q05	CLR = 1.6; df = 6; p = 0.95	CLR = 0.8; df = 3; p = 0.85
R102	-	R102Q04A	CLR = 2.4; df = 3; p = 0.49	CLR = 0.1; df = 1; p = 0.73
R104	R104Q01 & R104Q05	R104Q01	CLR = 0.5; df = 5; p = 0.99	CLR = 5.3; df = 3; p = 0.15
R111	-	R111Q06B	CLR = 9.0; df = 6; p = 0.18	CLR = 5.5; df = 3; p = 0.14
R219	R219Q01E & R219Q01T	R219Q01T	CLR = 3.0; df = 4; p = 0.56	CLR = 10.8; df = 2; p = 0.005
R220	R220Q04 & R220Q06 R220Q05 & R220Q06	R220Q05 R220Q06	CLR = 9.8; df = 7; p = 0.20	CLR = 3.0; df = 3; p = 0.39
R227	R227Q01 & R227Q02T R227Q03 & R227Q06	R227Q03 R227Q06	CLR = 7.1; df = 9; p = 0.62	CLR = 16.3; df = 5; p = 0.006

Note: Item R220Q02B was not administered among Danish students and is therefore not included in the analysis of R220

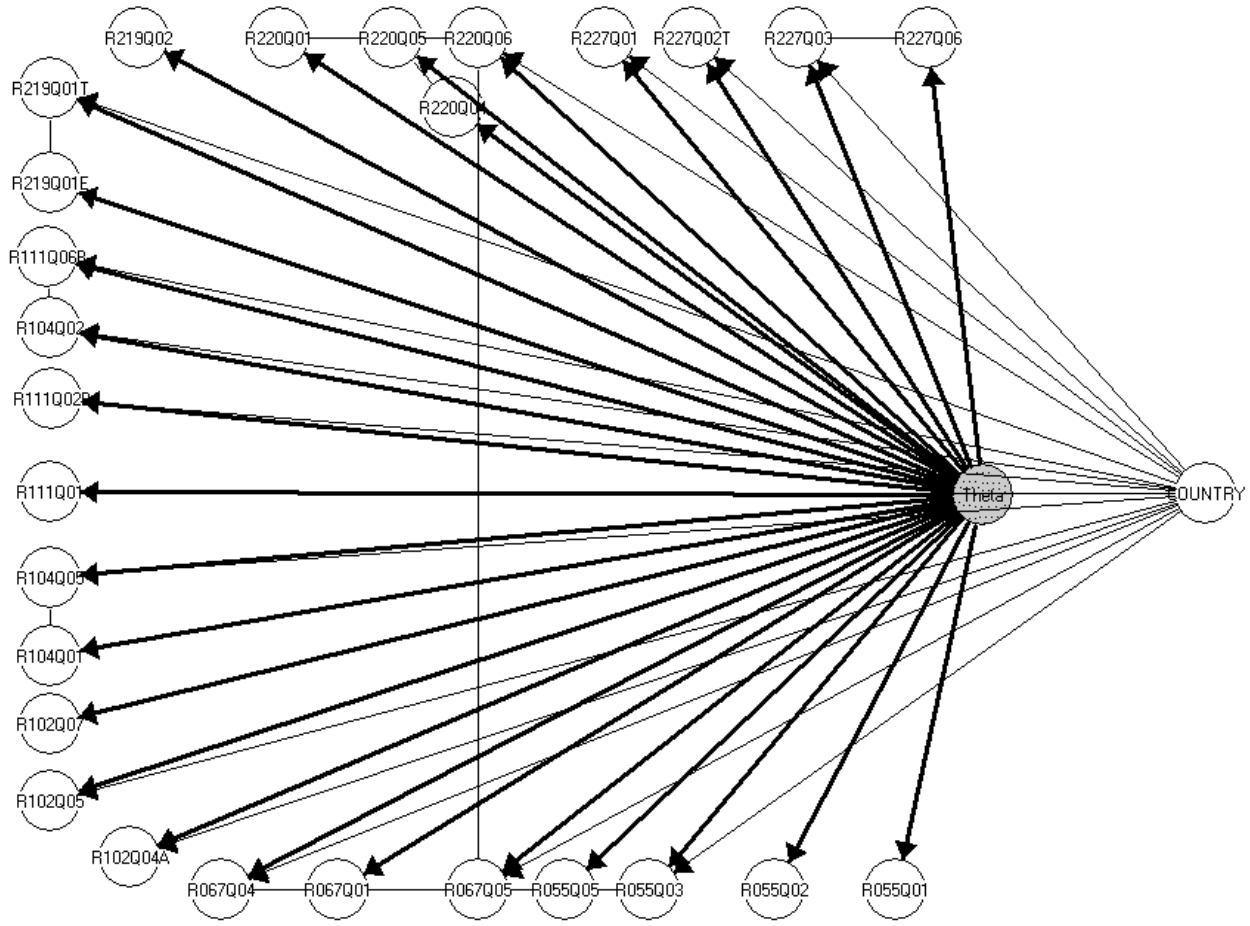


Fig. 6. Markov graph of a graphical loglinear Rasch model describing responses to 27 reading items in UK and Denmark

Let Y_1, Y_2, and Y_3 be responses to three items and let X be an exogenous variable. The total score is $S = Y_1+Y_2+Y_3$ and restscores are $R_{-1} = Y_2+Y_3$, $R_{-2} = Y_1+Y_3$, and $R_{-3} = Y_1+Y_2$. If the three items fit a Rasch model without DIF, the following hypotheses of conditional independence are all true,

$$Y_1 \perp Y_2 \mid R_{-1},\ Y_1 \perp Y_2 \mid R_{-2},\ Y_1 \perp Y_3 \mid R_{-1},\ Y_1 \perp Y_3 \mid R_{-3},\ Y_2 \perp Y_3 \mid R_{-2},\ Y_2 \perp Y_3 \mid R_{-3},\ Y_1 \perp X \mid S,\ Y_2 \perp X \mid S,\ Y_3 \perp X \mid S.$$

The modified item screening procedure tests these hypotheses for all triplets of items and collects the results in a graph where each item is represented by a node. If all hypotheses are accepted, item screening concludes that the three items could fit a Rasch model. In such cases, edges connecting all three items are added to the graph.

Assume next that a larger subset of items fits the Rasch model. Except for Type I errors, exploratory item screening will connect all the items in the set of items. Since cliques in graphs are maximal subsets of nodes that are completely connected, it follows that a set of items from a Rasch model will be contained in a clique. The natural next step is to check whether all items in a clique fit a Rasch model. This may not be so because of Type II errors during the analyses of triplets of items, but purification of items in cliques using standard methods for analysis by Rasch models should be able to finish the job in finite time.

In this case, this turned out to be true. The initial screening of item triplets defined a clique containing the following ten items: R055Q01, R055Q02, R067Q01, R102Q05, R102Q07, R104Q01, R111Q02B, R219Q01E, R219Q02, and R220Q04. The item analysis rejected the Rasch model for these items, following which it was relatively simple to finish purification by standard methods for Rasch analyses. The conditional likelihood ratio tests comparing item parameters among students with low and high scores and students from UK and Denmark accepted the model after elimination of R111Q02B and R219Q02 (CLR = 6.4, df = 7, p = 0.50, and CLR = 8.5, df = 7, p = 0.30 respectively), and item fit statistics did the same (Table 8). Finally, comparison of reading attainment in UK and Denmark based on these scores found no difference between the two countries (average scores were 5.7 in UK and 5.8 in Denmark).

7 DISCUSSION

To answer the question of the title first: no; PISA's measurement foundation is not solid. It is crumbling. Our analyses of data on reading from 2006 shows that responses to items violate all assumptions of the Rasch model on which they base their conclusions: items are locally dependent, item discriminations do not appear to be the same for all items, and, worst of all, there is extremely strong evidence of DIF relative to Country with a potentially strong effect on how countries are ranked. There is no avoiding the conclusion that PISA's scaling model cannot carry the burden of comparison of 56 countries and that the ranking published by PISA 2006 loses all credibility if nothing is done to correct the errors.

PISA's use of plausible values $\mathcal{P}$ also deserves comments. If the Rasch model fits and if the latent variable Θ has the conditional normal distribution assumed by PISA, then the distribution of $\mathcal{P}$ is the same as the distribution of Θ so that the ranking of countries by average plausible values is justified. If the model does not fit or if Θ has a non-normal distribution, then $\mathcal{P}$ is nothing but a random function depending on the total score S. If items fit an IRT model with monotonic

relationships between item scores and Θ *and if there had been no* DIF, then E(℘| Θ = θ) is an increasing function of θ so that ranking of countries by E(℘) should replicate the ranking of countries by E(Θ); but ranking by E(℘) is no better and probably more flawed than ranking by E(S), since the added random element during generation of plausible values means that persons with higher scores can lead to lower plausible values. There is, however, DIF, for which reason ranking by E(℘) is confounded just as ranking by E(S) is confounded.

On a more constructive note, our analyses of data on UK and Denmark suggest that the foundation could be fortified and that comparison could be made viable for at least some countries for two reasons. First, modelling by graphical loglinear Rasch models succeeded for the separate reading units. Despite the fact that modelling ultimately failed for the complete set of items, it suggests that DIF and local dependence could be adjusted for if attention was restricted to subsets of items. Second, purification actually identified eight items that fitted the Rasch model. Eight items is not much, but taken together with the success of GLLRM modelling of reading units, it is conceivable that it would be possible to add items to the subset of pure Rasch items and to adjust for the DIF and the local dependence that might turn up by GLLRM modelling. We do not expect this to be possible with data from all countries, but the fact that it was possible for UK and Denmark gives hope that it would also be possible, at least for other countries that are similar to UK and Denmark.

We developed the procedure for purification by exploratory item screening for this paper. It turned out to be relatively successful, because it did manage to identify a subset of Rasch items hidden within a very complex structure, but this does not mean that we are satisfied with the results as such or that we advocate an indiscriminate use of the procedure. First, eight dichotomous items do not provide reliable measurement, and standard errors of estimates of person parameters are very large. Second, content validity is dubious. Content validity requires that items cover all aspects of the trait that the educational test intends to measure. Since items are confidential, it is not possible to say much about this here, but the fact that six out of eight items are items tapping interpretation does not bode well for content validity. Finally, and perhaps most importantly, the question is whether it is meaningful to reduce comparisons of countries to a question of the average scores on a small set of items, when almost twice the number of items function differently. Does it make sense to say that reading attainment is the same in UK and Denmark because the average scores on a small subset of items are the same while scores on 15 items go both ways? We do not think so.

Table 8. Estimates of item parameters and item fit statistics based on Booklet 6 data on 20 PISA items.

				Item-restscore gamma								
item	thresholds	infit	p	observed	expected	p	chi	df	p	clr	df	p
R055Q01	-1.092	1.052	0.400	0.543	0.585	0.310	7.5	7	0.38	1.04	1	0.31
R055Q02	1.194	1.018	0.534	0.494	0.510	0.612	10.1	6	0.12	0.39	1	0.53
R067Q01	-1.294	0.929	0.293	0.633	0.595	0.372	5.5	7	0.60	1.06	1	0.30
R102Q05	1.043	0.965	0.236	0.539	0.512	0.367	6.2	6	0.40	3.22	1	0.07
R102Q07	-0.949	1.009	0.881	0.606	0.577	0.476	6.7	7	0.46	1.32	1	0.25
R104Q01	-1.051	0.942	0.338	0.640	0.582	0.162	6.5	7	0.48	2.32	1	0.13
R219Q01E	0.712	1.041	0.194	0.484	0.518	0.286	2.0	7	0.96	0.09	1	0.77
R220Q04	1.436	1.019	0.507	0.493	0.509	0.614	3.9	7	0.91	0.40	1	0.53

Another question is whether PISA's scaling model is also inadequate for mathematics and science. It is important to stress that we have only looked at data on reading and that our results do not automatically generalize to other areas. Following the discouraging results on reading, we cannot automatically trust PISA's results on mathematics and science, and we do suggest that it is better not to take PISA's results seriously until PISA publishes results supporting the Rasch model and the claim of no DIF among mathematics and science items. We could, of course, analyze data on these areas also, but the line must be drawn somewhere, and at the end of the day, the burden of proof that PISA's scaling model is adequate has to lie on the shoulders of PISA. In this paper, we have taken the first step with reading and have illustrated one way to do it, but the rest of the work must be up to PISA.

With regard to the ranking of countries, it would have been nice, had we been able to say that the true ranking of UK and Denmark is much higher than the ranking according to PISA, but this paper can say nothing of this. If item responses had been missing completely at random, if the subsample of students responding to Booklet 6 constituted a representative sample of students, and if there had been no DIF, then ranking of students based on the total scores over items that have been administered in all countries is valid under the majority of IRT models. Unfortunately, there is DIF, item responses are not missing at random, and the situation after the analyses is exactly as it was before the analyses. For these reasons, we do not have convincing information on the ranks of UK and Denmark or, for that matter, most other countries.

It is to be expected that our conclusions will generate counter arguments, and some of these are easily foreseen. A standard argument in such situations is to say that the claims made by the opponent "are based on misunderstandings related to the methodology underlying these international studies and a lack of research of the relevant technical documentation" (Adams (2003) replying to Prais (2003)). In response to this, we can say that we do understand Rasch model methodology and that we have indeed consulted the technical documentation from 2000, 2003, 2006, and 2009. If we have overlooked a part where PISA admits that the Rasch model does not fit and that there is DIF, and where they have also provided convincing evidence that their plausible values are robust to these departures from PISA's scaling model, then we will happily acknowledge this. However, PISA would need to tell us exactly where to look for it first. It has not been easy to find the documentation that shows that that they are aware of the problems, that they have attempted to solve the problems, what their solution was, and that their solution actually works.

Another argument could be that the Rasch model is an inappropriate model for this kind of data because the idea that all items have the same item discrimination is unrealistic, and that a better choice of a scaling model would be a 2-parameter IRT model. We do not share these reservations, nor do we expect this kind of argument from PISA since they do in fact use the Rasch model. If they did not believe in the adequacy of the Rasch model, they should, of course, have used another model instead. The problem remains, however, and has to be addressed during the test-of-fit of the model. We expect that our analysis of discrimination coefficients in Section 4 will suggest to some that the 2-parameter model had been a better choice of an initial scaling model than the Rasch model. Since evidence of item discrimination can emerge due to local dependence and DIF and because it actually proved to be possible to identify a subset of items that did not disagree with the Rasch model, we do not think that it is inconceivable that some kind of Rasch structure can be found. But the final word on this topic has not been said in this paper.

Finally, they may claim that they know about the problems, that the problems have been solved or that they their analyses show that the ranks provided by PISA are robust to the model errors. The truths of such claims are not supported by evidence in the technical reports and our results suggest that the ranking is far from robust. If they want to restore the credibility of their results, it is PISA's obligation to produce the evidence supporting their claims.

REFERENCES

Adams, R.J. (2003) Response to 'Cautions on OECD's Recent Educational Survey (PISA)'. *Oxford Review of Education*, **29**, 379-389

Adams, R.J., Wu, M.L., & Carstensen, C.H. (2007) Application of Multivariate Rasch models in International Large-Scale Educational Assessments. In M. Von Davier & C.H. Carstensen (eds). *Multivariate and Mixture Distribution Rasch Models*. New York: Springer.271-280

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69-81.

Brown, G., Micklewrigth, J., Schnepf, S.V. and Waldmann, R. (2007). International Surveys of educational achievement: how robust are the findings?. *J.R.Statist.Soc.* A, **170**, 623-646

Dorans, N.J. & Holland, P.W. (1993) DIF Detection and Description: Mantel-Haenszel and Standardization. In P.W. Holland & H. Wainer, H. (eds) *Differential item Functioning*. Hilsdale: Lawrence Erlbaum Associates.(pp 35-66)

Fischer, G.H. and Molenaar, I.W. (eds.) (1995) *Rasch models. Foundations, Recent developments, and*

applications. 293-306, New York:: Springer-Verlag.

Goldstein, H. (2004) International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education*, **11**, 319-330.

Goodman, L.A & Kruskal, W.H. (1954) Measures of Association for Cross Classifications. *J.Amer.Statist.Assoc.*,**49**, 732-764.

Holland, P.W. & Wainer, H. (editors) (1993) *Differential item Functioning*. Hilsdale: Lawrence Erlbaum Associates.

Hopmann, S.T., Brinek, G. & Retzl, M. (eds.) (2007) *PISA zufolge PISA . PISA According to PISA.* Wien: Lit Verlag

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, 49, 223-245.

Kelderman, H. (1989), Item bias detection using loglinear IRT. *Psychometrika*, 54, 681-697

Kirsch et.al. (2002) *Reading for change. Performance and Engagement across countries. Results from PISA 2000.* OECD

Kreiner, S. (2011) A note on item-restscore association in Rasch models. *Applied Psychological Measurement* (forthcoming)

Kreiner, S. & Christensen, K.B. (2002). Graphical Rasch models. In Mesbah, M., Cole, F.C. and Lee, M.T. (2002) *Statistical Methods for Quality of Life Studies.* Dordrecht: Kluwer Academic Publishers, p 187 – 203.

Kreiner, S. & Christensen, K.B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. Communications in Statistics. Theory and Methods, 33, 1239-1276.

Kreiner, S. & Christensen, K.B. (2006) Validity and objectivity in health related summated scales: analysis by graphical loglinear Rasch models. In von Davier, M. & Carstensen, C.H. (eds.) *Multivariate and Mixture Distribution Rasch models - Extensions and Applications.* New York, Springer Verlag, 329-346

Kreiner, S. & Christensen, K.B. (2011a) Exact evaluation of bias in Rasch model residuals. *Advances in Mathematics Research*, 12. (forthcoming)

Kreiner, S. & Christensen, K.B. (2011b) Item screening in graphical loglinear Rasch models. *Psychometrika* (forthcoming)

Lauritzen, S.L. (1996). *Graphical Models*. London. Clarendon Press.

OECD (2000) *Measuring Student Knowledge and Skills. The PISA 2000 Assessment of Reading, mathematical and Scientific literacy*. OECD, Paris. http://www.oecd.org/dataoecd/44/63/33692793.pdf

OECD (2006). *PISA 2006. Technical report*. OECD, Paris. http://www.oecd.org/dataoecd/0/47/42025182.pdf

OECD (2007). *PISA 2006. Science competencies for Tomorrow's world. Volume 1: Analysis*. OECD, Paris. http://www.oecd.org/dataoecd/30/17/39703267.pdf

Prais S.J. (2003) Cautions on OECD's Recent Educational Survey (PISA). *Oxford Review of Education*, **29**, 139-163

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Nielsen & Lydiche. Reprinted 1980 by MESA Press, Chicago

Schmitt, A.P. & Dorans, N.J. (1987) *Differential item functioning on the Scholastic Aptitude Test.* Research memorandum No. 87-1. Princeton NJ: Educational Testing Service

Van der Ark, L.A. & Bergsma, W.P. (2010) A Note on Stochastic Ordering of the Latent Trait Using the Sum of polytomous Item Scores. *Psychometrika*, 75, 272-279

Appendix A. Information on countries.

Table A provides information on average scores and the number of students with complete responses to twenty items.

Table A. Average total scores on twenty items in 56 countries.

	N	Mean	Std. Deviation	Std. Error
Azerbaijan	407	6.8501	4.59970	.22800
Argentina	257	10.1673	5.80941	.36238
Australia	1068	14.7584	5.71206	.17479
Austria	371	14.3181	5.54112	.28768
Belgium	653	15.0888	5.61014	.21954
Brazil	611	9.1751	5.71513	.23121
Bulgaria	328	10.4512	6.15081	.33962
Canada	1738	15.2048	5.69343	.13657
Chile	364	13.1016	5.31313	.27848
Chinese Taipei	668	14.9371	5.41298	.20943
Colombia	246	10.8821	5.18380	.33051
Croatia	400	13.4925	5.19011	.25951
Czech Republic	431	15.0278	6.18845	.29809
Denmark	357	14.5882	4.75125	.25146
Estonia	287	15.7944	5.07802	.29975
Finland	339	17.1150	4.39557	.23873
France	347	14.6023	5.62281	.30185
Germany	358	14.5698	5.56191	.29396
Greece	362	13.5994	5.77532	.30354
Hong Kong-China	345	16.1971	4.88727	.26312
Hungary	342	13.9444	5.47453	.29603
Iceland	289	13.9135	5.51036	.32414
Indonesia	732	8.7158	4.17616	.15436
Ireland	343	14.9446	5.49427	.29666
Israel	313	12.1693	6.58233	.37206
Italy	1611	13.8318	6.02394	.15008
Japan	441	14.2653	5.65403	.26924
Jordan	455	9.8176	4.74222	.22232
Korea	381	17.5249	5.05262	.25885
Kyrgyzstan	343	4.8688	4.21120	.22738
Latvia	357	14.4482	5.28095	.27950
Liechtenstein	27	13.7778	7.11625	1.36952
Lithuania	363	12.8843	5.40645	.28376

Table A (cont.). Average total scores on twenty items in 56 countries.

	N	Mean	Std. Deviation	Std. Error
Luxembourg	344	13.6948	5.68794	.30667
Macao-China	352	14.7983	4.96847	.26482
Mexico	2125	11.3280	5.06578	.10989
Montenegro	334	8.4401	4.78997	.26210
Netherlands	372	15.3333	5.35597	.27769
New Zealand	361	15.3546	5.81201	.30590
Norway	361	12.9086	6.23476	.32815
Poland	415	15.0120	5.44825	.26744
Portugal	390	13.4846	5.54936	.28100
Qatar	462	5.3874	4.87182	.22666
Romania	387	8.5375	4.86401	.24725
Russian Federation	392	13.0816	5.39854	.27267
Serbia	362	9.8978	5.30489	.27882
Slovak Republic	361	13.9584	5.71654	.30087
Slovenia	483	13.6149	5.45680	.24829
Spain	1476	13.6565	5.06894	.13194
Sweden	325	14.6677	5.62350	.31194
Switzerland	919	14.4559	5.44751	.17970
Thailand	452	10.8805	4.86436	.22880
Tunisia	304	9.3257	5.08760	.29179
Turkey	368	12.9647	5.49970	.28669
United Kingdom	1013	14.2270	5.66992	.17814
Uruguay	301	12.3754	6.12388	.35297
Total	28593	13.0213	5.97898	.03536

Appendix B

The output produced during the analysis of DIF includes country wise information on items that are significantly more difficult or easier in the country than expected by the Rasch model. Table B1 shows these items and reports the ranks of the country according to the score over easy and difficult items. The lists of easy and difficult items are not the final words on the relative difficulties of DIF items in different countries. The analysis failed to identify a subset of neutral anchor items that are equally difficulty in all countries for which reason all items must be regarded as DIF items. In addition to this, the ranks reported in the table are not the extreme ranks obtainable by examination of all subsets of items. Denmark, for instance, is no. 2 according to the score over items that are easy in Sweden, but no. 3 according to the items that are easy in Denmark.

Table B1. Country ranks defined by different item sets (R = rank)

Country	Difficult items	R	Easy items	R	Country	Difficult items	R	Easy items	R
Azerbaijan	CDEGNSWYZab	55	BFKLMTX	43	Korea	BCDEPVX	3	FGNYa	1
Argentina	MSWYZab	50	FGP	41	Kyrgyzstan	CKSTab	56	FGWXY	55
Australia	BFLMPVXa	20	CDKNSYZb	7	Latvia	CGNb	28	DEFa	3
Austria	CFGLP	34	BWXZb	6	Liechtenst.	FGL	48	KNZb	5
Belgium	FGNPY	27	CDKTVWXZab	5	Lithuania	BNYZb	41	CDEMVW	9
Brazil	CDKMNTWYZ	54	BFGPVa	44	Luxembourg	CFGLNP	42	TVZab	11
Bulgaria	CKPSVY	47	EFGLMN	44	Macao	BELX	48	FKMNSTWZ	5
Canada	ELMNVWXZ	29	BCDFGPSb	3	Mexico	DKMTWZa	47	FGLPb	31
Chile	DKLMTZa	44	CEFGPS	9	Montenegro	CLNPTYb	54	EKMZa	45
Taipei	BEPSXZ	40	FGKLMNTW	2	Netherlands	EFGMNP	30	KSYZab	2
Colombia	BDKMSTZab	50	EFGLP	19	N. Zealand	FMV	23	CKNSTb	2
Croatia	CVYb	45	DEFGKMSWZ	6	Norway	BCFGLV	47	KNPTZb	10
Czech Rep	FGPVa	21	CELTY	3	Poland	FKVYZb	30	BCDENTa	4
Denmark	CFGMVa	42	BKNSWXYZb	3	Portugal	KMNPYZ	41	CEFGLSVX	4
Estonia	FGXYZb	31	BCDKMPSWa	1	Qatar	CKSTb	55	FVXYZ	54
Finland	FGLPZ	4	DTVWXYab	1	Romania	NSYZb	53	CDELMV	42
France	GMP	40	BCDLSTVb	2	Russian Fed	LPZb	43	DEMNYa	11
Germany	FGLP	36	BKNVWXYZ	6	Serbia	CLNPTVXYb	52	DEFKMSZ	41
Greece	CDMSTWa	41	FGYZ	4	Slovak Rep.	Zab	38	EGPS	8
Hong Kong	BEGPXb	29	CDKMNSVWYZa	1	Slovenia	BFGM	46	CDELSXYZ	9
Hungary	BKLV	41	STXZ	13	Spain	FGPYZa	41	CEKMNTVXb	9
Iceland	CFVXZ	39	DKPSTYab	8	Sweden	FGVYa	35	KNTXZb	4
Indonesia	CDKMPSTYab	53	BFLNVWX	46	Switzerland	FGLPS	39	BDKMVXYZab	6
Ireland	LMX	33	BDES	3	Thailand	CEKTb	51	DFGLNPWZa	42
Israel	GTWX	45	FLP	28	Tunisia	BCEKMSTb	54	FLPVZa	43
Italy	DKMNSTWY	36	CEVXZa	15	Turkey	BCENWXZb	45	FGLMSTY	5
Japan	BCDEFGMVX	39	KLNPTYZb	4	UK	FPTVWX	36	BCKLSb	8
Jordan	CGKSTb	52	BEFPWXa	43	Uruguay	BTVWZa	42	EFP	10

Note: See Table for the definition of the item labels.

Appendix C. Assessment of ranking error.

Let Y_{cvi} be the score on item i by person v from country c (c = 1,..,C; v = 1,..,N_c; i = 1,..,I) and let A be the indices of a subset of items $A \subset \{1,..,I\}$. This appendix is concerned with errors when countries are ranked according to the average country scores over all items $S_c = \frac{1}{N_c}\sum_{i=1}^{I} Y_{cvi}$ and over items in A $T_c = \frac{1}{N_c}\sum_{i\in A} Y_{cvi}$.

We assume that the latent trait parameters have conditional normal distributions with means ξ_c and standard deviations σ_c and that consistent estimates of item and population parameters are available. The estimates of the item parameters can be used for calculation of the conditional distribution of scores given the latent trait. Finally, data from the different countries also provides estimates of the distributions of the person scores, and therefore also of $E(S_c)$ and $SD(S_c)$.

Under the Rasch model, country ranks according to $(S_1,..,S_C)$ and $(T_1,..,T_C)$ are expected to be similar, but ranking errors will occur depending on the number of items and on sample sizes in different countries: the smaller the sample size and the smaller the number of items, the larger the ranking error.

To evaluate the size of the ranking error for a country, we generate Nsim random sets of country means and for each set count the number of countries with a higher country mean in the sample. Generating random country means can be done in three different ways of which only one depends on the assumption that the latent trait variable has a conditionally normal distribution.

1) If sample sizes are large enough, it follows that the average country scores have approximately normal distributions with expected value equal to $E(S_c)$ and standard errors equal to $SE(S_c) = SD(S_c)/\sqrt{N_c}$. The inexpensive way to generate random average country scores is to sample from these distributions.
2) A slightly more expensive approach is to generate random person scores from the estimated score distribution and then calculate the average country scores used for ranking.
3) The expensive approach is to generate random values of the latent trait first and then generate random scores from the conditional score distribution.

Our analyses of ranking errors for UK and Denmark when S_c was used as criterion followed the inexpensive approach generating 100,000 random samples of country means.

During our analyses of ranking errors when T_c was use for ranking, we assumed that items fitted a Rasch model and used a three-step procedure combining the first and the third approaches.

First, item parameter estimates from the analysis of the complete set of items were used for calculation of the conditional distribution of the subscore given the latent variable. Second, a random sample of 10,000 person parameters was generated for each country. The population parameters used during this step were those obtained during the analysis of the complete set of items. Third, random scores were generated from the distribution of the subscore estimates for each of the random persons, providing Monte Carlo estimates of $E(S_c)$ and $SD(S_c)$. Finally, the inexpensive approach generating 100,000 random samples of country means was used to estimate the distribution of the ranks of UK and Denmark by the subset of items.