PSYCHOMETRIKA—VOL. 79, NO. 2, 210–231
APRIL 2014
DOI: 10.1007/S11336-013-9347-Z

ANALYSES OF MODEL FIT AND ROBUSTNESS. A NEW LOOK AT THE PISA SCALING MODEL UNDERLYING RANKING OF COUNTRIES ACCORDING TO READING LITERACY

SVEND KREINER

UNIVERSITY OF COPENHAGEN

KARL BANG CHRISTENSEN

UNIVERSITY OF COPENHAGEN

This paper addresses methodological issues that concern the scaling model used in the international comparison of student attainment in the Programme for International Student Attainment (PISA), specifically with reference to whether PISA's ranking of countries is confounded by model misfit and differential item functioning (DIF). To determine this, we reanalyzed the publicly accessible data on reading skills from the 2006 PISA survey. We also examined whether the ranking of countries is robust in relation to the errors of the scaling model. This was done by studying invariance across subscales, and by comparing ranks based on the scaling model and ranks based on models where some of the flaws of PISA's scaling model are taken into account. Our analyses provide strong evidence of misfit of the PISA scaling model and very strong evidence of DIF. These findings do not support the claims that the country rankings reported by PISA are robust.

Key words: differential item functioning, ranking, robustness, educational testing, programme for international student assessment, PISA, Rasch models, reading literacy.

1. Introduction

The Programme for International Student Assessment (PISA) is an ambitious large-scale attempt to measure and compare proficiency in reading, mathematics, and science in a large number of countries. The first PISA survey was launched in 2000, with follow-up surveys in 2003, 2006, 2009 and 2012. Many concerns regarding the comparability of test results from different countries have been raised, in particular, difficulties producing items that are culturally and linguistically neutral. Prais (2003), Goldstein (2004), Brown et al. (2007), and Hopmann, Brinek, and Retzl (2007) discuss some of these issues.

The assumption underlying PISA is that *true* country ranks according to reading literacy can be defined in meaningful ways by the ranking of countries according to the means of a latent variable of a unidimensional IRT model. PISA generates so-called "plausible values" with the same distribution as the latent trait variable and reports country ranks defined by the averages of plausible values that are *estimates* of the true ranks. The results reported by PISA show clearly that not all observed differences among countries are statistically significant. This fact appears to be ignored by the public, the press, and by the politicians responsible for educational policy. Apparently, the key feature to them of the PISA studies is the ranking system, regardless of whether or not the differences in ranks reflect substantive differences in aptitude between countries. Therefore, it is crucial to assess the systematic and unsystematic errors associated with the estimates of the ranks. This is the main focus of this paper.

Requests for reprints should be sent to Svend Kreiner, Department of Biostatistics, University of Copenhagen, Oster Farimagsgade 5, B, PO Box 2029, 1014 Copenhagen K, Denmark. E-mail: s.kreiner@biostat.ku.dk

Close inspection reveals that PISA is founded on the assumption that item responses fit a Rasch model (Fischer and Molenaar, 1995). PISA has taken pains to check the fit of item responses to the Rasch model and, in particular, the assumption that item difficulties were homogeneous across countries. Purification by elimination of items that did not fit the Rasch model was also carried out during the initial item analysis. Chapters 9 and 12 of the technical report OECD (2006) describe the methods used during model checking.

However, the technical reports (e.g., OECD, 2000, 2006, 2007, and 2009) leave many questions unanswered. The data collected by PISA are freely accessible and easy to download and analyze. Therefore, in the first part of this paper, we test whether the Rasch model is adequate for responses to PISA items and, in particular, whether there is evidence of differential item functioning (DIF) between countries. We use the same type of model-testing techniques as described in Chapters 9 and 12 of the OECD technical report (2006). Our results show that the fit of the model is inadequate and that there is very strong evidence of country related DIF.

The second part of the paper examines the robustness of ranking of countries using two different methods: (a) by analysis of the degree to which country ranks are invariant across subscales of items and (b) by comparison of country ranks based on the Rasch model with country ranks based on models where some of the flaws of the Rasch model have been taken into account.

1.1. The DIF Issue

It is well known that educational tests that contain items which exhibit DIF "may have reduced validity for between-group comparisons" (Dorans and Holland, 1993; Schmitt and Dorans, 1987). The issue as to which attempts PISA has made to ensure that there is no DIF in terms of country is fundamental, as the main purpose of PISA is to compare literacy across countries.

PISA acknowledges the importance of DIF stating that "The interpretation of a scale can be severely biased by unstable item characteristics from one country to the next" (OECD, 2006, p. 86), and that the "consistency of item parameter estimates across countries was of particular interest" (OECD, 2006, p. 147). According to the same technical report (p. 104), "Out of 179 mathematics, reading and science items, 28 items were omitted in a total of 38 occurrences for the computation of national scores" due to DIF. Furthermore, Kirsch et al. (2002) state 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." These statements suggest that PISA has attempted to deal with DIF by elimination of items. However, Adams (2003) reports that "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," suggesting that elimination only addressed DIF due to very specific problems. PISA's analysis of DIF is described in Chapter 9 of OECD (2006) and briefly summarized in Section 3 of this paper. There is no information in any technical report about DIF analyses after elimination of items. A question to PISA concerning this issue remains unanswered. We, therefore, assume that no such analysis was carried out and that it, therefore, is impossible to see whether DIF was dealt with adequately.

DIF could also have been dealt with by item-splitting, where DIF items are replaced by sets of virtual items, one for each country, with different item difficulties. Adams, Wu, and Carstensen (2007, p. 274) state when discussing PISA methodology 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)." Such analyses were, however, not attempted (personal communication from Ray Adams). However, we briefly pursue this possibility in Sections 4.1 and 5.4 of this paper where we show that splitting items as proposed by Adams et al. (2007) has an effect on the ranking of countries.

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, 11, 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.

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

FIGURE 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.)

2. Data on Reading in PISA 2006

The reading inventory of PISA 2006 contains 28 items organized in eight testlets called "reading units," each containing three to five items with a common text. The reading units are included in 14 booklets, but only one (Booklet 6) contains all reading units and six booklets (Booklets 1, 3, 4, 5, 8, 10) have no reading units at all. Thus, roughly half of the student participating in the PISA 2006 survey did not respond to any reading items. In spite of this, all students were assigned reading scores (so-called plausible values). Exactly how these scores were calculated is one of the unanswered questions, but the brief discussion of plausible values in Chapter 9 of OECD (2006) suggests that they may be random numbers drawn from the conditional distribution of the latent reading ability given scores on math and science items and a number of person covariates. Plausible values are briefly described in Section 3 of this paper, but we propose that readers interested in PISA's plausible values consult Chapter 9 of OECD (2006) and the references found in this document.

Table 1 contains a list of the reading units and the booklets in which they appear, and Figure 1 presents an overview of the data on reading in the PISA 2006 survey. The reading units are partitioned into two sets with 14 items in each. Set 1 consists of R055, R104, R111, and R227. Set 1 appears in Booklets 6, 9, 11, and 13. Set 2, consisting of R067, R102, R219, and R220, is included in Booklets 2, 6, 7, 12.

TABLE 2.
Reading items, PISA 2006.

Item	Item type	Maximum item score	Included in all countries	Item set 1[1]	Item set 2[1]
R055Q01	Interpreting	1		(+)	
R055Q02	Reflecting	1	+	+	
R055Q03	Interpreting	1	+	+	
R055Q05	Interpreting	1	+	+	
R067Q01	Interpreting	1	+		+
R067Q04	Reflecting	2	+		+
R067Q05	Reflecting	2	+		+
R102Q04A	Interpreting	1			(+)
R102Q05	Interpreting	1			(+)
R102Q07	Interpreting	1			(+)
R104Q01	Information	1	+	+	
R104Q02	Information	1	+	+	
R104Q05	Information	2	+	+	
R111Q01	Interpreting	1	+	+	
R111Q02B	Reflecting	2		(+)	
R111Q06B	Reflecting	2	+	+	
R219Q01E	Interpreting	1			(+)
R219Q01T	Information	1			(+)
R219Q02	Reflecting	1	+		+
R220Q01	Information	1	+		+
R220Q02B	Interpreting	1			(+)
R220Q04	Interpreting	1	+		+
R220Q05	Interpreting	1	+		+
R220Q06	Interpreting	1	+		+
R227Q01	Interpreting	1	+	+	
R227Q02T	Information	2	+	+	
R227Q03	Reflecting	1	+	+	
R227Q06	Information	1	+	+	

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

Table 2 contains information about the 28 reading items, 22 dichotomous items, and six partial credit items with scores 0, 1, and 2. PISA distinguishes between three types of reading items: Retrieving of information from texts (7 items), interpretation (14 items), and reflection (7 items). The texts and items of the reading units are not available to the public, although examples of reading units that have not been used for assessment of students can be found in OECD (2000).

The PISA 2006 data can be downloaded at http://pisa2006.acer.edu.au/downloads.php. These data contain responses to reading items from 56 countries. In the dataset, there are eight items without responses in some countries, but 20 items have responses in all countries. In addition to items that are completely missing in some countries, a number of students have missing responses to some items so that the average number of manifest item responses per student is 8.3.

The missing responses do not appear to be completely randomly missing, because the frequencies of students with at least one missing item response is very different across countries. 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 20 items. The technical report offers no explanation for these differences.

The 20 items included in all countries comprise 15 dichotomous items and five partial credit items scored 0, 1, and 2. The total score thus ranges between zero and 25.

3. PISA Methodology

Information on PISA's analyses of educational test data can be found in the OECD Technical Report (2006), specifically Chapter 9 "Scaling PISA cognitive data," and Chapter 11 "Scaling outcomes." The following section provides a brief summary of what is to be found in these chapters, although the interested reader should consult the OECD technical reports and the references in this report for a complete account of PISA's item analysis.

3.1. The Scaling Model

The PISA scaling model contains items $(Y_1, \ldots, Y_{28})$, a unidimensional latent variable Θ, the variable country C, and other exogenous variables X including gender, age, and grade of student, education and occupation of parents, type and size of schools, class sizes, and many other variables. The measurement component of the scaling model is an ordinary Rasch model (the six polytomous items are modelled using the partial credit model). Thus, the scaling model assumes that items are locally independent and without DIF. Finally, Θ is assumed to be conditionally normal with means ξ_{cx} that depend on C and X, $\xi_{cx} = E(\Theta|C = c, X = x)$. Appendix 2 of OECD provides a list of exogenous variables. It is not clear whether all variables are used in all countries.

3.2. Assessing the Fit of Item Responses to Rasch Models

According to PISA, "particular attention was paid to the fit of the items to the scaling model, item discrimination, and item-by-country interaction" during data analysis (OECD, 2006, p. 147). The methods used for this are briefly described and illustrated in the technical report (OECD, 2006, pp. 147–152). The issue of item fit was addressed in three ways: by calculation of Infit item fit statistics (Smith, 2004); by calculation of discrimination coefficients; and by informal comparisons 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 summary results were assembled in tables in national reports that are not available to the public, and examples of the content are shown in the technical report. These examples indicate that Infit test statistics and discrimination coefficients have been calculated for the subsample of 15,000 students. However, the actual results are not found in the technical report.

3.2.1. Infit Test Statistics Let Y_{vi} denote the observed item score for person v on item i and let E_{vi} denote the expected score. The Infit test statistic (Smith, 2004) is the sum of squared *un-standardized* residuals, divided by the sum of the variances of the residuals,

$$Infit_i = \frac{\sum_v (Y_{vi} - E_{vi})^2}{\sum_v Var(Y_{vi} - E_{vi})}. \tag{1}$$

The expected item score and the variance of the residuals can easily be derived from the Rasch model. In practice, item and person parameters are unknown and are replaced by estimates of the parameters during calculation of expected item scores. In the PISA analysis, there is no formal assessment of the significance of these test statistics. The way in which they are reported suggests that values outside the interval [0.7, 1.3] are regarded as evidence of inadequate fit of the item to the Rasch model.

3.2.2. Discrimination Coefficients PISA uses the point-biserial correlation coefficient (Glass and Hopkins, 1995) as a measure of discrimination and requires point-biserial discriminations to be above 0.25 for all items. The issue of discrimination could also be addressed during Rasch analyses, but there is no evidence in the technical reports of PISA doing this.

3.3. Country Comparisons

PISA uses a multistep procedure for the analyses of country differences.

Step 1. National item calibration: MML estimates of the item parameters are calculated separately for each country. Despite the fact that PISA expects Θ to depend on a number of exogenous conditioning variables, the MML estimates are based on the assumption that students have been sampled from a normal distribution (OECD, 2006, p. 146).

Step 2. International item calibration: MML estimates of the item parameters of the Rasch model are calculated on a random subset of 15,000 students from 30 OECD countries. These estimates also assume that the 15,000 students are sampled from a normal distribution, even though it is a fundamental assumption made by PISA that Θ depends on the country.

Step 3. The outcomes of the national calibrations are used to make a decision about how to treat each item in each country. This is based on item fit statistics and on evaluation of DIF by informal comparisons of MML estimates from the international and national item calibrations (OECD, 2006, pp. 147–152). Items with unsatisfactory psychometric properties are either deleted altogether or treated as not-administered in particular countries.

Step 4. PISA's population model is a conditional multivariate normal distribution describing the conditional distribution of literacy in reading, math and science, given a number of so-called conditioning variables including gender, school, occupational status, and educational levels of parents, and many other variables. PISA estimated the population model in each country using MML item parameter estimates from the international calibration as known parameters (OECD, 2006, p. 155). Appendix 2 of OECD (2006) provides a complete list of conditioning variables. Whether they are all used in all countries and whether the fit of the population model to data was assessed is an open question because there is no evidence in the technical reports that the fit was assessed.

Step 5. During this step, PISA uses the MML estimates of item parameter from the international calibration and the estimates of the population models in each country to generate so-called plausible values representing literacy in reading, math, and science.

Step 6. Calculations of plausible values: Plausible values $\tilde{\Theta}$ are random numbers drawn from the posterior distribution of Θ, given (Y, C, X). Calculation of plausible values is particularly simple under the Rasch model because plausible values can be drawn from the conditional distribution of Θ, given S, C, X, instead of the (more complicated) conditional distribution of Θ, given Y, C, X. Estimates of item and population parameters are needed to determine the posterior distribution. PISA uses the MML estimates of item parameter from the international calibration and the estimates of the population models in each country to generate the plausible values (OECD, 2006, pp. 155–156).

Step 7. Analysis of variance: PISA finally compares countries using a one-way ANOVA of the plausible values $\tilde{\Theta}$ (OECD, 2009). Cluster sampling is taken into account; however, it is not clear as to whether the added error due to imputation of missing responses and generation of plausible values has been taken into account.

Using plausible values may appear to be an attractive way to conduct a latent-structure analysis based on Rasch models because the conditional distribution of $\tilde{\Theta}|C$ (that is $\tilde{\Theta}|S, C, X$ marginalized over the total score S and X) is the same as the distribution of $\Theta|C$.

The advantage of analysis by plausible values is less obvious when one recalls that generation of plausible values require estimates $\hat{\xi}_{cx}$ of the conditional means of Θ, given C and X. To

see why, let $\tilde{\xi}_{cx}$ be the average of the plausible values for $C = c$ and $X = x$. These averages are supposed to estimate $\xi_{cx} = E(\Theta|C, X)$, but in fact, they are estimates of $\hat{\xi}_{cx}$ rather than ξ_{cx} because plausible values were generated under a model where Θ is Gaussian with the mean equal to $\hat{\xi}_{cx}$. If we assume that $\tilde{\xi}_{cx}$ is an unbiased estimate of $\hat{\xi}_{cx}$, it follows that $\tilde{\xi}_{cx}$ is also an unbiased estimator of ξ_{cx}. The standard error of the $\tilde{\xi}_{cx}$ as an estimate of ξ_{cx} has to be larger than the standard error of $\hat{\xi}_{cx}$ because the standard error of $\tilde{\xi}_{cx}$ as an estimate of $\hat{\xi}_{cx}$ is added to the standard error of $\hat{\xi}_{cx}$. A second reason as to why the advantage of analysis by plausible values is less than obvious is that it is difficult to see why plausible values are needed when estimates of $E(\Theta|C, X)$ already exist. Finally, a third reason to suspect that plausible values may be less than plausible is the assumption that the distribution of Θ is conditionally Gaussian. If this assumption is false, it follows that the density of distribution of the plausible values $\tilde{\Theta}$ is not the same as the density of distribution of Θ; and thus, there is no reason to be interested in the distribution of $\tilde{\Theta}$ at all.

It is sometimes argued that plausible values are better suited for secondary analyses of factors related to Θ than estimates of person parameters. This may be true, if programs for marginal item analysis and latent regression are unavailable and if we can trust in the claims on behalf of the prior distribution of Θ. If programs for latent regression are available and if we can believe in the prior distribution of Θ, it seems obvious that analysis by these programs are preferable to secondary analyses of plausible values using standard routines for statistical analysis. If there, on the other hand, is reason to be concerned about the prior distribution of Θ, we prefer secondary analyses of either person scores or person parameter estimates to analysis of plausible values.

4. Item Analysis of Booklet 6 Data on Reading

Booklet 6 is the only booklet with all the 28 reading items, but item responses were only recorded for 20 items in all countries. For this reason and because misfit of the Rasch model to a subset of items implies misfit to the complete set of items, we restricted the analyses of the adequacy of the Rasch model to an analysis of the fit of the Rasch model to Booklet 6 data on the 20 items with responses from all countries.

A total of 30,605 students were exposed to Booklet 6, but only 28,593 students had complete responses to the 20 reading items with responses from all countries. Since we use conditional maximum likelihood (CML) estimates of item parameters and conditional likelihood ratio (CLR) tests to test the fit of the model, our analysis makes no assumption on how persons are distributed and how they are sampled. For this reason and because a sample consisting of 28,593 respondents is an unusually large sample, we restricted the analysis to a check of the Rasch model among students with complete responses. Appendix A provides country information on the number of students and the average scores on the 20 items.

We attempted to replicate PISA's analysis. We did this by calculating the Infit test statistics and discrimination coefficients and comparing item parameter estimates from different countries to item parameters estimated in the complete dataset for analysis of DIF. However, there were also differences in the analysis procedures, as described below.

First, PISA uses marginal maximum likelihood (MML) estimates of item and population parameters, assuming that Θ has a conditional normal distribution given exogenous variables. Since it is known that the MML approach produces biased item parameters if the distribution of Θ is misspecified (Adams et al., 1997), we prefer CML estimates and CLR tests to methods based on MML estimates.

Second, we test the hypothesis that item parameters are the same for students with low scores and students with high scores using Andersen's (1973) conditional likelihood ratio (CLR) test. There is no evidence that PISA calculated a similar overall test-of-fit. In addition to the CLR test

for the complete Booklet 6 dataset, we also calculated the CLR test for the separate countries. The results of these tests can be seen in Appendix A.

Third, the significance of departure of observed fit statistics from expected values under the Rasch model is assessed for all fit statistics and discrimination coefficients.

Fourth, the Infit test statistic used compares observed item scores to expected values from the *conditional* distribution of item responses given the total score, thus avoiding the bias inherent in the traditional Infit test statistics. The asymptotic distribution of this test statistic is easily derived (Kreiner and Christensen, 2011).

Fifth, we used Goodman and Kruskal's (1954) γ to measure the degree of association between item scores and restscores on other items, because it is more appropriate for analysis of association among ordinal categorical variables. We calculated the expected value under the Rasch model and assessed significance of the difference between the observed and expected coefficients (Kreiner, 2011a, 2011b). To further strengthen the assessment of the claim that all items discriminate in the same way, we calculated and assessed the significance of Molenaar's U (Molenaar, 1983).

Sixth, the analysis of DIF in the PISA analysis relies on an informal comparison of item parameter estimates in different countries. We supplement this by a formal test of DIF using the Andersen (1973) CLR test.

Finally, two different tests are used to assess DIF for individual items: (i) a χ^2 test of conditional independence of item and country, given the total score on all items, and (ii) a likelihood ratio test of the Rasch model against a loglinear Rasch model, where item parameters for the item are different in different countries (Kelderman 1984, 1989). The χ^2 tests of conditional independence of item score and country, given the total score on all items, are tests in large sparse tables. To avoid the well-known problems with asymptotic p-values for such tests, we used Monte Carlo estimates of exact conditional p-values based on 1000 random tables for each test (Kreiner, 1987).

4.1. Results

The evidence against the Rasch model is overwhelming. The CLR tests 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 it rejects the hypothesis that item parameters are the same in all countries (CLR = 27,389.0; df = 1320, $p < 0.00005$). The item fit statistics in Table 3 tell the same story. All but three items are rejected by all item fit statistics, and all items are rejected by the tests for DIF. Some of the evidence against the Rasch model in Table 3 and Appendix A may be due to Type I errors, and it is likely that the Type I error rate is inflated since many of the analyses are based on the Rasch model. Some of this evidence may be spurious, but attempts to purify the set of items failed to identify a subset of items without DIF (results not shown).

Following the rejection of the Rasch model, we attempted to model uniform DIF and local dependence (LD) among items from the same testlets by adding two-factor loglinear interaction terms to the model (Kelderman, 1984), that is, models where the strength of the local dependence of items is assumed to be independent of Country and independent of Θ. These attempts also failed to produce a satisfactory model; however, models defined in this way were better according to the Bayesian information criterion (BIC). The initial BIC for the Rasch model was 540293. Stepwise inclusion of interaction terms between items and Country until the BIC increased resulted in a model with BIC equal to 526188. In this model, all but four items have DIF. Subsequently, we added loglinear interaction terms describing LD between items within testlets until, once again, BIC began to increase. This resulted in a very complicated model with BIC equal to 518950. Table 4 summarizes the DIF and LD among items according to the two models. Despite the complexity of these models, item fit statistics also reject them (results not shown).

TABLE 3.
Item fit statistics assessing the fit of the Rasch model to responses to 20 PISA items.

	Item Infit		Item-restscore gamma			Molenaar's U		χ^2 test of no DIF			Conditional likelihood ratio test of no DIF		
	Infit	p	Obs.	Exp.	P	u	p	χ^2	df	P	CLR	df	P
R055Q02	0.968	<0.0001	0.554	0.522	<0.0001	3.93	<0.0001	896.1	684	0.000	680.2	55	<0.0001
R055Q03	0.894	<0.0001	0.633	0.537	<0.0001	11.03	<0.0001	1099.3	684	0.000	1213.2	55	<0.0001
R055Q05	0.822	<0.0001	0.719	0.559	<0.0001	16.28	<0.0001	1230.9	684	0.000	630.5	55	<.0001
R067Q01	0.978	0.055	0.601	0.590	0.154	−0.42	0.68	1208.6	684	0.000	891.3	55	<0.0001
R067Q04	1.242	<0.0001	0.457	0.578	<0.0001	−15.23	<0.0001	2644.4	1316	0.000	3230.5	110	<0.0001
R067Q05	1.233	<0.0001	0.544	0.643	<0.0001	−11.34	<0.0001	2913.7	1316	0.000	3515.3	110	<0.0001
R104Q01	0.899	<0.0001	0.681	0.572	<0.0001	10.20	<0.0001	1373.6	684	0.000	885.8	55	<0.0001
R104Q02	1.166	<0.0001	0.325	0.512	<0.0001	−21.02	<0.0001	1335.4	684	0.000	1284.7	55	<0.0001
R104Q05	0.998	0.782	0.563	0.531	<0.0001	2.99	0.0028	2997.6	1103	0.000	2222.3	110	<0.0001
R111Q01	0.971	<0.0001	0.586	0.545	<0.0001	2.98	0.0029	1049.4	684	0.000	594.9	55	<0.0001
R111Q06B	1.123	<0.0001	0.565	0.620	<0.0001	−7.05	<0.0001	1883.8	1262	0.000	2231.8	110	<0.0001
R219Q02	0.891	<0.0001	0.663	0.567	<0.0001	9.49	<0.0001	1488.2	684	0.000	1104.4	55	<0.0001
R220Q01	0.871	<0.0001	0.653	0.513	<0.0001	13.96	<0.0001	1108.5	632	0.002	868.7	55	<0.0001
R220Q04	1.003	0.545	0.539	0.532	0.253	−2.18	0.029	932.5	684	0.000	717.3	55	<0.0001
R220Q05	0.923	<0.0001	0.670	0.563	<0.0001	9.87	<0.0001	994.8	684	0.000	368.7	55	<0.0001
R220Q06	1.055	<0.0001	0.498	0.537	<0.0001	−6.70	<0.0001	1044.4	684	0.000	1075.3	55	<0.0001
R227Q01	1.132	<0.0001	0.398	0.524	<0.0001	−18.78	<0.0001	1183.3	684	0.000	1494.5	55	<0.0001
R227Q02T	1.099	<0.0001	0.501	0.559	<0.0001	−9.31	<0.0001	2856.8	1316	0.000	3359.9	110	<0.0001
R227Q03	0.913	<0.0001	0.615	0.530	<0.0001	9.30	<0.0001	1216.9	684	0.000	656.5	55	<0.0001
R227Q06	0.862	<0.0001	0.679	0.548	<0.0001	14.28	<0.0001	1354.7	684	0.000	1500.0	55	<0.0001

Note: The p-values for the χ^2 test of no DIF are Monte Carlo estimates of exact conditional p-values (Kreiner, 1987).

TABLE 4.
Overview of DIF and local dependence among items according to the loglinear Rasch models described in Section 4.1.

Item	Country DIF	Local dependence
R055Q02	yes	R055Q03 and R055Q05
R055Q03	yes	R055Q02 and R055Q05
R055Q05	no	R055Q02 and R055Q03
R067Q01	yes	R067Q04 and R067Q05
R067Q04	yes	R067Q01 and R067Q05
R067Q05	yes	R067Q01 and R067Q04
R104Q01	no	R104Q02 and R104Q05
R104Q02	yes	R104Q01
R104Q05	yes	R104Q01
R111Q01	yes	R111Q06B
R111Q06B	yes	R111Q01
R219Q02	yes	
R220Q01	no	R220Q04, R220Q05 and R220Q06
R220Q04	yes	R220Q01 and R220Q06
R220Q05	no	R220Q01, R220Q04 and R220Q06
R220Q06	yes	R220Q01 and R220Q05
R227Q01	yes	R227Q02T
R227Q02T	yes	R227Q01, R227Q03 and R227Q06
R227Q03	yes	R227Q02T and R227Q06
R227Q06	yes	R227Q02T and R227Q03

To assess the impact of DIF on the total scores, Kreiner and Christensen (2007) propose a DIF equating procedure where scores are adjusted to make them comparable with the scores of a reference group. The adjusted scores for comparison with a reference country can be found in Appendix A. These show the impact of DIF to be changes in country means ranging from −3.6 to 2.6. Beyond statistical significance, these are substantial changes, considering that the possible score ranges from 0 to 25, and that the observed country means range from 4.87 to 17.52.

4.2. Analysis of Data from Other Booklets

In this paper, we have focused on data from Booklet 6 because this was the only booklet with all items. However, we also examined the adequacy of the Rasch model on data from Booklets 2, 7, 9, 11, 12, and 13, with 14 reading items in all. In addition to confirming the inadequacy of the Rasch model and producing strong evidence of DIF relative to country, these analyses also disclosed strong evidence of DIF relative to booklet. The booklet effect is probably due to the location of the reading units within the booklets. Each reading unit appears in four booklets in one of the following four locations: First 30-minute cluster, second 30-minute cluster, third 30-minute cluster, and fourth 30-minute cluster.

The overall test statistics calculated during these analyses are presented in Appendix B. Item fit statistics and other details are not reported here, but are available on request from the authors.

5. Country Ranking

5.1. Introduction

We distinguish between systematic and unsystematic ranking errors. The unsystematic errors of the estimated ranks depend on sample sizes and on the true differences among countries.

Under the Rasch model and under the assumption that all students had responded to all items, country ranks could be consistently estimated using the total sum score because the total score is sufficient for the person parameters of the model. Ranking by the average scores under more general IRT models would probably also be correct because of the monotone relationship between item responses and the latent variable in these models.

Since some students have not responded to all items, a more complicated latent structure analysis or an analysis of plausible values, as used by PISA, is needed. Using plausible values adds extra random error compared to the random error of a latent structure analysis, but this will not be problematic when sample sizes are large if the scaling model fits the data.

Unfortunately, PISA's scaling model is inadequate. The effect of using plausible values generated by a flawed model is unknown. Under the Rasch model, plausible values are generated from the conditional distribution of the person parameter Θ given the score S over all items. Since $E(S|\Theta = \theta)$ is an increasing function of θ under all unidimensional IRT models that satisfy the requirement of monotonicity, local independence and no DIF (Rosenbaum, 1989) and since $E(\Theta|S = s)$ is an increasing function of the score whether or not the conditional distribution of Θ given S is correctly specified; the lack of fit might not lead to systematic errors as long as items are locally independent and there is no DIF. There is DIF, however, and the amount of DIF evidence is so great that the risk of systematic ranking errors cannot be dismissed.

In this section, we therefore address the robustness of PISA's ranking. We do this in two ways: (a) by examination of the degree to which rankings of countries are invariant across item subsets, and (b) by comparison of rankings based on person parameter estimates from loglinear Rasch models.

The assessment of the invariance of the ranking will also be based on the responses to the 20 items that have been administered in all countries among the subset of 28,593 students with responses to all 20 items. The analysis therefore excludes students that were not exposed to Booklet 6, and Booklet 6 students with one or more missing responses to the 20 items. For this reason and because we cannot assume that item responses are missing at random and because of the DIF relative to booklet, we do not expect that a ranking of countries based on such a dataset would be able estimate the true ranking of countries without bias even if the Rasch model had been adequate and without country DIF. Our intention is, however, less ambitious than that. In this section, we only investigate the invariance of country rankings by data collected in the way that the data on the 20 items in Booklet 6 were collected. If the country ranks in this dataset were found to be invariant, we would still have to be concerned about invariance in the data collected by other booklets, but our belief in the invariance of country ranks based on the complete dataset would be strengthened. Since there was evidence against invariance, it follows that ranking by data on the other booklets also is suspect and that ranking by the complete PISA dataset cannot be expected to be invariant.

5.2. Unsystematic Random Error Under the Rasch Model

The random error of rankings by average scores from Rasch models can be examined by parametric bootstrapping. Methods for that purpose are briefly described in Appendix C. For Denmark, the closest thing to a 95 % confidence interval for the rank based on the Booklet 6 data on the 20 items administered in all countries is [10, 24]. Thus, the unsystematic error is considerable. Recall, however that the ranking error would have been smaller had the sample been larger. Our calculations show that the confidence interval would be [14, 20] if 2000 students had responded to the 20 items in all countries. This probably better reflects what the random error in PISA would have been if the Rasch model had been adequate, since the total number of manifest responses to reading items under this setup is similar to the situation for PISA where the average number of item responses is less than 9.

TABLE 5.
Ranking of 5 countries by 8 different subscores based on Booklet 6 data. The numbers in parentheses indicate the number of items included in the subscore.

Observed Booklet 6 Rank	Country	Items no.			Item set		Item type		
		0xx (6)	1xx (5)	2xx (9)	1 (12)	2 (8)	Information (6)	Interpretation (8)	Reflection (6)
7	Canada	3^{**}	12	21^{*}	8	5^{*}	25^{***}	18^{+}	2^{***}
16	France	17	28^{+}	14	23	18	16	15	23
17	Denmark	36^{**}	20	5^{*}	7^{+}	32^{*}	6^{*}	8	37^{**}
22	Japan	40^{**}	11	9^{+}	13	36^{*}	8^{*}	30	28
23	UK	14^{*}	23	30^{+}	22	30^{+}	23	26	18

Notes: The observed Booklet 6 rank is the rank according to the 20 items administered to all students. The degree to which rankings that are improbable under the Rasch model is assessed by one-sided p-values and indicated in the following way: $^{+}$: $0.05 < p \leq 0.10$; *: $0.01 < p \leq 0.05$; **: $0.001 < p \leq 0.01$; ***: $p \leq 0.001$.

5.3. *Invariance*

To many, invariance is regarded as a fundamental property of Rasch models. Invariance implies that country ranks—except for statistical errors—will not depend on which items are used provided they come from the same Rasch model. In this section, we examine the degree to which PISA's reading items live up to this challenge by comparison of country rankings of five countries (Canada, France, Denmark, Japan, and UK) by different subsets of items.

The item overview in Table 2 defines three different ways to partition the items. First, the reading units have code numbers starting with 0, 1, or 2. What these numbers refer to is not public knowledge, but this is inconsequential for the check of the invariance. Second, the reading units are divided into two subsets with 14 items, only one of which is administered to the majority of the students. And finally, PISA distinguishes between three types of items—information, interpretation, and reflection.

Table 5 shows country ranks by subscores defined by these criteria. The number of items collected in the subscores is limited because subscores only summarize responses to the items that were administered in all countries. Comparison of rankings by different subscores, therefore, has to allow for considerable random error because of the low number of items. Letting R_c denote the observed rank of country c, the methods described in Appendix C can be used to calculate the probabilities $P(R_c \leq r)$ and $P(R_c \geq r)$. The results are shown in Table 5. The most extreme cases are those found for Canada, where the probability of a rank equal to 25 or worse on items measuring information retrieval is equal to 0.00022, whereas the probability that the rank is equal to 1 or 2 on items relating to reflection is equal to 0.00008. For Canada, Denmark, Japan, and the UK, subscale ranks are unlikely under the Rasch model, but subscale ranks for France do not differ from expected values under the Rasch model.

The p-values used to assess the discrepancy between the observed subscore ranks and the ranks expected by the Rasch model are one-sided. The indications of too extreme ranks in Table 5 therefore refer to an 80 % confidence region for the ranks. Since half the ranks lie outside this confidence region, we conclude that Table 5 does not support claims of invariance of PISA's ranking.

Next, we compared the ranks for 1000 random subsets of 14 items according to Booklet 6 data to the ranks of countries, and according to simulated data from a Rasch model. The choice of 14 items was motivated by the fact that PISA administers 14 items to the majority of students exposed to booklets with reading items. The data were generated using a Rasch model with the same sample sizes and the same item and population parameters as (estimated) in the Booklet

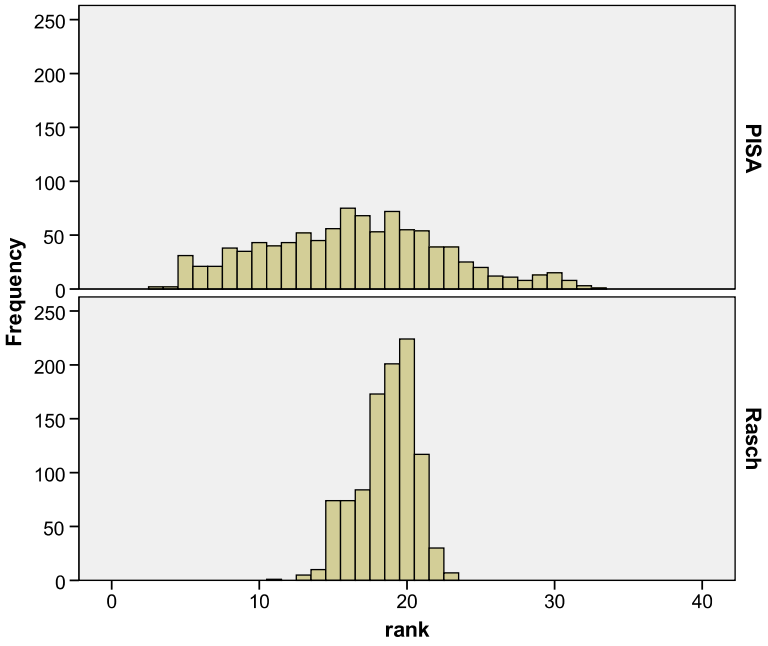


FIGURE 2.
Distribution of the rank of Denmark. The *top panel* shows the rank in 1000 random subsets of items consisting of 14 items from the PISA data. The *bottom panel* shows the variation of ranks 1000 datasets simulated using a Rasch model with the item and population parameters estimated from PISA's data.

TABLE 6.
Variation of ranks by 1000 random subscores with 14 items. Results are shown both for data from PISA and for data simulated from a Rasch model with the same parameters as those estimated in PISA's data.

Country	Booklet 6 data				Rasch		
	Rank[1]	Mean	s.d.	Range	Mean	s.d.	Range
Canada	7	8.2	3.4	3–21	8.4	1.3	6–13
France	16	16.4	3.8	5–27	16.4	1.9	10–23
Denmark	17	16.4	6.1	3–33	18.6	2.0	11–23
Japan	22	21.5	5.0	6–36	22.2	1.3	18–25
UK	23	22.1	2.8	15–30	25.3	0.6	23–27

[1]: The rank reported here is the rank according to the 20 items administered to all students.

6 dataset. Figure 2 shows results for Denmark, and Table 6 summarizes the results for the five countries discussed above.

The results in Figure 2 and Table 6 show that the variation of ranks in PISA's data is far beyond the variation found in data from the Rasch model. In all countries, the standard deviations and ranges of the ranks are far larger in the data collected by PISA than in the data from the Rasch model. The differences are largest in Denmark and Japan but are also pronounced in the other countries.

The total number of possible subscores is much larger than 1000, and the ranges in Table 6 therefore underestimate the true ranges of ranks across all possible subscores. To get an idea about the range of ranks across all possible subscores, we used observed and expected item scores in different countries obtained during calculation of the conditional likelihood ratio tests to select item subsets that would provide extreme ranks. In Denmark, for instance, item scores

TABLE 7.
Ranks for five countries under three different models.

Country	Rasch model	GLLRM with DIF	GLLRM with DIF and LD
Canada	7	20	21
France	16	8	9
Denmark	17	13	14
Japan	22	25	20
UK	23	27	26

of R055Q02, R104Q01, R111Q01, R219Q02, R220Q05, R220Q06, R227Q01, R227Q02T, and R227Q06 are significantly higher than expected while item scores of the items R055Q03, R067Q04, R067Q05, R104Q05, R220Q04, and R227Q03 are significantly lower than expected. Ranking by these item-subsets places Denmark as no. 3, and no. 42, respectively. Similar estimates of the extreme ranges for the other countries are as follows: Canada 3-29, France 2-40, Japan 4-39, and the UK 8-36. Similar results for the other countries that took part in the PISA 2006 survey can be found in Kreiner (2011b).

5.4. Ranking by Models with DIF and Local Dependence

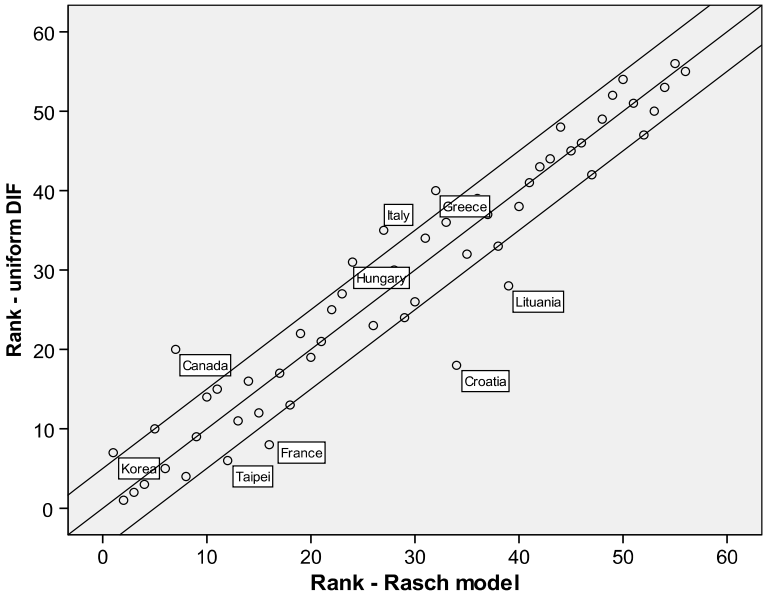
Another way to assess the robustness of the ranks based on the Rasch model is to compare them to ranks based on a more adequate model. It has not been possible to identify a completely adequate IRT model, but Section 4.1 described two loglinear Rasch models (one with DIF and one with DIF and local dependence) that fit the data better. Person parameters were estimated under these models and the ranks defined by these estimates compared to the ranks by the Rasch model. Table 7 shows the results for the five countries discussed above, and Figure 3 shows the results for all countries. The average difference between the ranks by the Rasch model and the ranks by the DIF and DIF+LD models are 3.5 and 4.2, respectively, but the discrepancies between the ranks by the Rasch models and the ranks by the better models are substantial for several countries including Canada and France. The most noticeable example is Croatia moving 23 steps upward when ranked by the DIF+LD model, but it is also of interest that Korea drops from no. 1 to no. 8 under the DIF+LD model. The analysis also shows that LD has a noticeable effect on the ranks. Croatia is once again the most extreme case, where the difference in rank under the two models is equal to 8.

6. Discussion

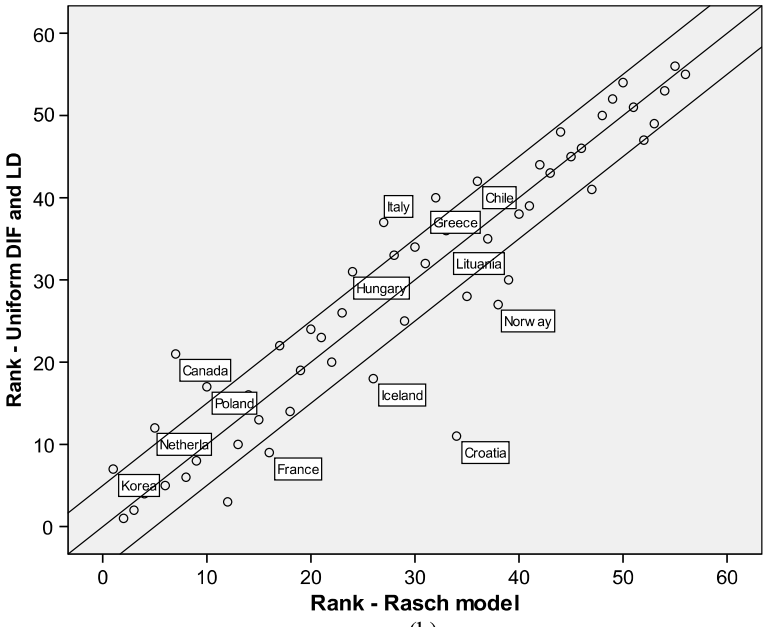
In this paper, we have examined the fit of the Rasch model to responses to the items PISA uses to measure reading literacy, and we have developed and applied methods for assessment of systematic and unsystematic ranking errors. There are two conclusions to be drawn from this study. The first is irrefutable and the other is open for discussion.

6.1. Model Fit and DIF

The first concerns the fit of items to PISA's scaling model. Despite the claims that "particular attention was paid to the fit of the items to the scaling model," evidence against the Rasch model is overwhelming. In particular, we note that within the spectrum of the Rasch model, there is strong evidence of DIF, and we have not been able to find an item subset fitting the Rasch model without DIF. Granted, of course, that statistical models never fit perfectly; and tests of fit in large sample studies such as PISA will always ultimately yield evidence against the model. For



(a)



(b)

FIGURE 3.
Rasch model ranks plotted against ranks by loglinear Rasch models with (**a**) DIF and (**b**) local dependence. Country names are displayed if the difference in ranks under the two models is larger than 5.

this reason, we also assessed the fit of the Rasch model separately in each country. As seen in Appendix A, the model was rejected in all countries except Lichtenstein, where the number of students is very small. Although some of these findings may be Type I errors (indeed, the Type error I rate may be inflated since analyses are based on a flawed model), the evidence against the Rasch model is irrefutable.

PISA's reading test has testlet structure; even so, the analyses reported by PISA do not use a testlet model. Fitting Rasch models to each testlet disclosed evidence of local dependence within testlets (results not shown). This means that a simple testlet model, where the local dependence within testlets can be described by a random effect, cannot fit the data. For this reason, we attempted to incorporate both DIF and testlet structure by adding loglinear item-country interaction terms and within-testlet loglinear interaction terms to the Rasch model. The end result was a loglinear Rasch model, which was better than the pure Rasch model but nevertheless also inadequate. It is for this reason important to point out that our analyses were inconclusive and that some of the evidence suggesting DIF from the perspective of both the pure Rasch model and our loglinear Rasch model may be due to other model errors. Having admitted that, we must also emphasize that the evidence of DIF is so strong that it is unrealistic to expect that a different IRT model can make all evidence of DIF disappear. For instance, the local testlet dependence in our loglinear Rasch model did not reduce the evidence of DIF at all.

Note, finally, that the analyses in this paper are complete-case analyses. If the results of our analyses had supported the Rasch model, it would not automatically have meant that the Rasch model fitted the complete dataset. However, our analysis showed the Rasch model to be inadequate for the complete set of Booklet 6 responses, and from this it follows that the Rasch model is inadequate for the complete dataset.

6.2. Country Ranks

The second conclusion concerns the degree to which country ranks are robust to the model errors. We have looked into this question in two ways: by analysis of invariance of rankings across different subscales, where we take the random error connected with ranking by small subsets of items into account; and by comparison of the Rasch model's ranking of countries with ranking of countries by loglinear Rasch models, where DIF and local dependence to some extent have been taken into account. It is our conclusion that the results do not support claims that the ranking of countries by PISA is viable. First, the variation of ranks by different subsets of items is above and beyond what it is possible to explain as random variation. And second, the country ranks produced by the Rasch model and the country ranks produced by other (better) models are different. It can be argued, however, that our conclusion is subjective and is dependent on our interpretation of the results. We have, however, documented what we have found and leave the reader to make his or her own decision as to the interpretation of these results.

As to the ranking of countries, it should be noted that item responses are not missing at random and that ranking by the average of scores over the 20 items for students who responded to all items may be biased relative to the true ranking. This is not important in terms of our conclusions drawn in this paper. For we do not claim that we can rank countries, but only that rankings are not invariant and that models with better fit to the data yield country ranks different from those produced by the Rasch model.

Finally, concerning our investigation of the invariance of ranking by different subsets of items, we also point out that PISA has published results on a similar analysis of invariance for the science items (Adams et al., 2010). During these analyses, subsets of items were defined by items judged by each country to be of highest priority for inclusion; and rankings were defined by an approximate average percent-correct methodology. The number of science items defining subscales during PISA's analysis of invariance range from 2 to 81.

PISA concluded that the results "show a remarkable consistency in the country rank orderings across different sets of countries' preferred items" but also noted "that these results should be considered as indicative only" and that "a more rigorous analysis would be to use the full PISA scaling methodology and use a wider variety of approaches for selecting item subsets" (Adams et al., 2010, p. 12). This is precisely what we have done for the Booklet 6 data on reading, and our results show a remarkable inconsistency in the country ranks.

6.3. The Missing Scaling Model

Unfortunately, our attempts to find an adequate scaling model for the PISA data on reading failed. This does not mean that our loglinear Rasch models are flawed to the same extent as PISA's Rasch model, and we therefore expect the rankings by our models to be closer to the truth than the rankings by PISA. Remember also that the ranking of countries by PISA rests on the assumption that it is both meaningful and possible to rank countries according to the country average of a unidimensional latent variable in an IRT model. The ranking therefore depends on PISA being able to produce an adequate model. We do not believe that this is attainable, although PISA is more than welcome to prove us wrong; and, there are, of course, obvious ways to proceed from the best fitting loglinear Rasch model that we derived. The addition of item discrimination parameters depending on reading units/testlets is one obvious possibility. Developing IRT models for separate countries before attempting to consolidate into one model is another. Some of this is already under way, but the responsibility for finding a viable model ultimately lies with PISA, if they want to continue to claim that they can rank countries in a meaningful and reliable way.

6.4. The Distribution of Reading Ability According to PISA

The analysis used by PISA for computing country ranks is based on MML estimates of item parameters, where it is assumed that the latent variable is normally distributed. The 56 countries in the PISA study are very different with widely varied educational policies. For this reason, it is not likely that the latent variable is normally distributed. The distributions of the person parameter estimates based on the Rasch model are negatively skewed in all countries, and thus we would expect the MML parameter estimates to be biased to a certain degree. Our analyses used CML estimates that are known to be consistent and do not rely on distributional assumptions.

The skewed person parameter distribution raises questions concerning the degree to which it is appropriate to rank countries by the averages of the person parameters. Furthermore, if some countries put a lot of effort in raising low scores, while others focus on the high scores, a ranking of countries based on means may not be meaningful. Note, however, that the skewed distribution of the person parameters in different countries does not have to disagree with PISA's assumption that reading ability is *conditionally* normal, given appropriate conditioning variables. Nonetheless, a closer look at this assumption is definitely warranted.

6.5. Mathematics and Science

A question that we have not addressed at all is whether PISA's scaling model is also inadequate for mathematics and science. It is important to stress that our results cannot automatically be generalized to other areas, and we do not imply that the claims of Adams et al. (2010) are false. However, considering the discouraging results on reading literacy, we cannot automatically trust PISA's results on mathematics and science, and we suggest that they should not be considered to be well-founded until PISA publishes results supporting the Rasch model and the claim of no DIF.

6.6. Counter-Arguments

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, in response to Prais, 2003). To this charge, we can only convey that we quite thoroughly understand Rasch model methodology and that we have indeed consulted the technical documentation from 2000, 2003, 2006, and 2009. If we have overlooked some part where PISA admits that the Rasch

model does not fit and that there is DIF, and where they provide convincing evidence that their plausible values are robust in these departures from PISA's scaling model, then we will happily acknowledge any misconception. However, PISA would have to direct us exactly to where to locate this evidence. Apart from the report by Adams et al. (2010), we have not been able to find documentation indicating that they are aware of the problems and that they have attempted to solve them.

6.7. Methodological Issues

Our paper addresses two general methodological issues.

The first is about the importance of the choice of the statistical model. It is not unusual to experience arguments claiming that the specific choice of the statistical model does not matter and that the results using different statistical models are always similar. We do not subscribe to this point of view and think that the comparison of ranks by different models supports our reluctance to embrace such claims. The model does matter.

The second is the problem of model fitting in large sample studies. We expect it will be pointed out that models never fit in large sample studies and that our evidence of the misfit of the Rasch model is therefore to be expected, and thus of no interest. We do not accept this point of view, because it implies that we should always collect a lot of data to avoid the trouble of testing and correcting statistical models. Our perspective is exactly the opposite. In large sample studies, the chances of identifying errors that can easily be corrected are much better than in small sample studies, and it is easier to assess the effect of the model errors. We think that our paper makes this point.

Appendix A. Information on Countries

Table A.1 provides information on (a) average scores and the number of students with complete responses to 20 items, (b) DIF equated scores with Azerbaijan as reference country (Kreiner and Christensen, 2007), and (c) overall tests of fit of the Rasch model in 56 countries.

Appendix B. Analyses of Data from All Booklets with Reading Items

In addition to Booklet 6 with 28 reading items, reading items can also be found in Booklets 2, 7, 9, 11, 12, and 13. Each of these booklets contained 14 items. Booklets 9, 11, and 13 had items from reading units R055, R104, and R111. We refer the these booklets together with Booklet 6 that also had these reading units as Booklet set 1. Booklet set 2 consists of Booklets 2, 6, 7, and 12 with reading units R067, R102, R219, and R220.

Table B.1 shows the overall CLR tests of the Rasch model for the two different booklet sets as a whole and for the different booklets. In addition to the CLR tests not DIF relative to country, CLR tests also provided evidence of DIF relative to booklets (Booklet set 1: CLR = 1669.0, df = 42, $p < 0.00005$; Booklet set 2: CLR = 3260.3, df = 27 $p < 0.00005$).

Additional information on these analyses is available from the authors on request.

Appendix C. Assessment of Ranking Error

Let Y_{cvi} be the score on item i by person v from country c ($c = 1, \ldots, C$; $v = 1, \ldots, N_c$; $i = 1, \ldots, I$) and let A be the indices of a subset of items $A \subset \{1, \ldots, I\}$. The total score on all

TABLE A.1.
Average total and DIF equated scores on 20 items in 56 countries and conditional likelihood ratio (CLR) tests of the Rasch model comparing item parameters estimated for student with raw scores below and above the median raw score in the country.

	N	Mean	Std. Dev.	Std. Error	Equated mean	Average bias[1]	CLR[2]	*p*
Azerbaijan	407	6.85	4.60	0.23	6.85	0.00	105.29	<0.0001
Argentina	257	10.17	5.81	0.36	9.99	0.18	81.09	<0.0001
Australia	1068	14.76	5.71	0.17	15.77	−1.01	208.33	<0.0001
Austria	371	14.32	5.54	0.29	15.74	−1.43	122.82	<0.0001
Belgium	653	15.09	5.61	0.22	17.59	−2.50	100.74	<0.0001
Brazil	611	9.18	5.72	0.23	7.48	1.73	134.25	<0.0001
Bulgaria	328	10.45	6.15	0.34	11.51	−1.06	134.79	<0.0001
Canada	1738	15.20	5.69	0.14	15.88	−0.67	244.96	<0.0001
Chile	364	13.10	5.31	0.29	11.12	1.99	113.44	<0.0001
Chinese Taipei	668	14.94	5.41	0.21	17.70	−2.76	118.81	<0.0001
Colombia	246	10.88	5.18	0.33	9.60	1.29	59.50	0.0001
Croatia	400	13.49	5.19	0.26	17.17	−3.63	69.47	<0.0001
Czech Republic	431	15.03	6.18	0.30	17.30	−2.27	120.55	<0.0001
Denmark	357	14.59	4.75	0.25	16.33	−1.74	56.29	0.0002
Estonia	287	15.79	5.08	0.30	17.85	−2.06	71.92	<0.0001
Finland	339	17.12	4.40	0.24	20.71	−3.59	70.38	<0.0001
France	347	14.60	5.62	0.30	17.11	−2.51	95.40	<0.0001
Germany	358	14.57	5.56	0.29	15.51	−0.94	93.29	<0.0001
Greece	362	13.60	5.78	0.30	10.99	2.61	63.35	<0.0001
Hong Kong-China	345	16.20	4.89	0.26	18.45	−2.25	65.08	<0.0001
Hungary	342	13.94	5.47	0.30	15.77	−1.82	114.25	<0.0001
Iceland	289	13.91	5.51	0.32	15.94	−2.02	103.65	<0.0001
Indonesia	732	8.72	4.18	0.15	8.63	0.08	175.65	<0.0001
Ireland	343	14.94	5.49	0.30	16.56	−1.61	134.54	<0.0001
Israel	313	12.17	6.58	0.37	11.53	0.64	175.94	<0.0001
Italy	1611	13.83	6.02	0.15	13.42	0.42	396.41	<0.0001
Japan	441	14.27	5.65	0.26	15.63	−1.36	82.06	<0.0001
Jordan	455	9.82	4.74	0.22	10.37	−0.55	91.18	<0.0001
Korea	381	17.52	5.05	0.26	17.48	0.04	67.13	<0.0001
Kyrgyzstan	343	4.87	4.21	0.23	5.77	−0.90	173.45	<0.0001
Latvia	357	14.45	5.28	0.28	16.45	−2.01	92.39	<0.0001
Liechtenstein	27	13.78	7.12	1.36952	13.27	0.50	26.19	0.34
Lithuania	363	12.88	5.41	0.28376	15.26	−2.37	76.93	<0.0001
Luxembourg	344	13.69	5.69	0.30667	15.45	−1.76	65.67	<0.0001
Macao-China	352	14.80	4.97	0.26482	17.37	−2.57	78.47	<0.0001
Mexico	2125	11.33	5.07	0.10989	10.48	0.84	447.24	<0.0001
Montenegro	334	8.44	4.79	0.26210	8.76	−0.32	86.83	<0.0001
Netherlands	372	15.33	5.36	0.27769	16.49	−1.15	94.28	<0.0001
New Zealand	361	15.35	5.81	0.30590	17.26	−1.91	79.46	<0.0001
Norway	361	12.91	6.23	0.32815	14.95	−2.04	95.95	<0.0001
Poland	415	15.01	5.45	0.26744	16.94	−1.93	108.84	<0.0001
Portugal	390	13.48	5.55	0.28100	13.60	−0.11	117.08	<0.0001
Qatar	462	5.39	4.87	0.22666	5.47	−0.08	309.30	<0.0001
Romania	387	8.54	4.86	0.24725	9.66	−1.12	77.68	<0.0001
Russian Federation	392	13.08	5.40	0.27267	15.40	−2.32	63.83	<0.0001
Serbia	362	9.90	5.30	0.27882	11.69	−1.79	63.15	<0.0001
Slovak Republic	361	13.96	5.72	0.30087	15.27	−1.31	83.29	<0.0001

TABLE A.1.
(*Continued*)

	N	Mean	Std. Dev.	Std. Error	Equated mean	Average bias[1]	CLR[2]	*p*
Slovenia	483	13.61	5.46	0.24829	14.69	−1.07	111.88	<0.0001
Spain	1476	13.66	5.07	0.13194	14.64	−0.99	366.67	<0.0001
Sweden	325	14.67	5.62	0.31194	16.33	−1.66	106.66	<0.0001
Switzerland	919	14.46	5.45	0.17970	15.42	−0.97	159.48	<0.0001
Thailand	452	10.88	4.86	0.22880	12.48	−1.60	102.64	<0.0001
Tunisia	304	9.33	5.09	0.29179	8.35	0.98	47.46	0.0029
Turkey	368	12.96	5.50	0.28669	13.97	−1.00	115.78	<0.0001
United Kingdom	1013	14.23	5.67	0.17814	14.84	−0.61	195.91	<0.0001
Uruguay	301	12.38	6.12	0.35297	11.86	0.52	99.49	<0.0001

Notes: (1) Bias = observed − equated scores. (2) The degrees for the conditional likelihood ratio test are equal to 24.

TABLE B.1.
Overall fit statistics for Booklets sets 1 and 2 and for Booklets 2,7,9,11,12,13.

Booklets	*n*	Low and high score groups			Country DIF		
		CLR	df	*p*	CLR	df	*p*
Booklet set 1	111904	11193.2	14	0	53308.7	770	0
9	23594	3203.2	14	0	11917.0	770	0
11	28752	3222.9	14	0	14143.6	770	0
13	29717	2123.3	14	0	14934.3	770	0
Booklet set 2	113471	5523.0	9	0	39731.7	495	0
2	25395	1011.5	9	0	8866.8	495	0
7	29897	1782.4	9	0	11204.1	495	0
12	29652	969.6	9	0	12342.3	495	0

Note: CLR tests based on data on students with complete responses on all items administered in all countries.

items is $S_{cv} = \sum_{i=1}^{I} Y_{cvi}$ and the subscore over items in A is $T_{cv} = \sum_{i \in A} Y_{cvi}$. This Appendix is concerned with errors when countries are ranked according to averages $S_c = \frac{1}{N_c} \sum_v S_{cv}$ and $T_c = \frac{1}{N_c} \sum_v T_{cv}$.

We assume that item responses fit a Rasch model with a latent variable Θ. The distribution of Θ may be nonparametric or parametric. In the nonparametric case, the population parameters of interest are the score probabilities $P(S_{cv} = s)$ and $P(T_{cv} = t)$. The marginal distribution of T_{cv} is given by $P(T_c = t) = \sum_s P(T_{cv} = t | S_{cv} = s) P(S_{cv} = s)$. Under the Rasch model, $P(T_{cv} = t | S_{cv} = s)$ depends on item parameters, but not on Θ. Under such a model, it is consequently easy to calculate estimates of the subscore probabilities $P(T_{cv} = t)$ in the nonparametric case if consistent estimates of the item parameters and the score probabilities $P(S_{cv} = s)$ are available.

In the parametric case, we assume that the latent variables are Gaussian normal with means ξ_c and standard deviations σ_c. Given these distributions, Monte Carlo methods provide simple estimates of the distributions of S_{cv} and T_{cv} based on estimates of item parameters together with estimates of ξ_c and σ_c.

The country ranks according to $(S_1, \ldots, S_C)$ and $(T_1, \ldots, T_C)$ are expected to be similar under the Rasch model, 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. And, of course, the ranking errors also depend on ξ_c and σ_c. The results reported in this paper are derived under the parametric model with Monte Carlo estimates of the distributions of S_{cv} and T_{cv} based on Monte Carlo samples of 10,000 random students from each country.

To estimate the distribution of the country ranks based on country averages S_c and/or T_c, we generated 100,000 random values of S_c and/or T_c and for each set ranked the countries according to these values. Given these estimates, it is easy to find both confidence intervals and probabilities of extreme rankings for the countries.

References

Adams, R.J. (2003). Response to 'Cautions on OECD's recent educational survey (PISA)'. *Oxford Review of Education*, *29*, 379–389. Note: Publications from PISA can be found at http://www.oecd.org/pisa/pisaproducts/.

Adams, R., Berezner, A., & Jakubowski, M. (2010). *Analysis of PISA 2006 preferred items ranking using the percent-correct method*. Paris: OECD. http://www.oecd.org/pisa/pisaproducts/pisa2006/44919855.pdf.

Adams, R.J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

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* (pp. 271–280). New York: Springer.

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

Brown, G., Micklewrigth, J., Schnepf, S.V., & Waldmann, R. (2007). International surveys of educational achievement: how robust are the findings? *Journal of the Royal Statistical Society. Series A. General*, *170*, 623–646.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hilsdale: Lawrence Erlbaum Associates.

Fischer, G.H. & Molenaar, I.W. (Eds.) (1995). *Rasch models—foundations, recent developments, and applications*. Berlin: Springer.

Glass, G.V., & Hopkins, K.D. (1995). In *Statistical methods in education and psychology*. Boston: Allyn & Bacon.

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. *Journal of the American Statistical Association*, *49*, 732–764.

Hopmann, S.T., Brinek, G., & Retzl, M. (Eds.) (2007). *PISA zufolge PISA. PISA according to PISA*. Wien: Lit Verlag. http://www.univie.ac.at/pisaaccordingtopisa/pisazufolgepisa.pdf.

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, I., de Jng, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change. performance and engagement across countries. results from PISA 2000*. Paris: OECD.

Kreiner, S. (1987). Analysis of multidimensional contingency tables by exact conditional tests: techniques and strategies. *Scandinavian Journal of Theoretical Statistics*, *14*, 97–112.

Kreiner, S. (2011a). A note on item-restscore association in Rasch models. *Applied Psychological Measurement*, *35*, 557–561.

Kreiner, S. (2011b). Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. Research report 11/1, Dept. of Biostatistics, University of Copenhagen. https://ifsv.sund.ku.dk/biostat/biostat_annualreport/images/c/ca/ResearchReport-2011-1.pdf.

Kreiner, S., & Christensen, K.B. (2007). Validity and objectivity in health-related scales: analysis by graphical loglinear Rasch models. In M. Von Davier & C.H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 271–280). New York: Springer.

Kreiner, S., & Christensen, K.B. (2011). Exact evaluation of bias in Rasch model residuals. *Advances in Mathematics Research*, *12*, 19–40.

Molenaar, I.V. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, *48*, 49–72.

OECD (2000). *Measuring student knowledge and skills. the PISA 2000 assessment of reading, mathematical and scientific literacy*. Paris: OECD. http://www.oecd.org/dataoecd/44/63/33692793.pdf.

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

OECD (2007). *PISA 2006. Volume 2: data*. Paris: OECD.

OECD (2009). *PISA data analysis manual: SPSS* (2nd ed.). Paris: OECD. http://www.oecd-ilibrary.org/education/pisa-data-analysis-manual-spss-second-edition_9789264056275-en.

Prais, S.J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education*, *29*, 139–163.

Rosenbaum, P. (1989). Criterion-related construct validity. *Psychometrika*, *54*, 625–633.
Smith, R.M. (2004). Fit analysis in latent trait measurement models. In E.V. Smith & R.M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove: JAM Press.
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.

Manuscript Received: 8 DEC 2011
Final Version Received: 30 JAN 2013
Published Online Date: 14 JUN 2013