How robust are cross-country comparisons of PISA scores to the scaling model used?

John Jerrim (UCL Institute of Education and Education Datalab)

Philip Parker (Australian Catholic University, Institute for Positive Psychology and Education)

Alvaro Choi (Institut d'Economia de Barcelona)

Anna Katyn Chmielewski (University of Toronto)

Christine Sälzer (Technical University of Munich, School of Education)

Nikki Shure (UCL Institute of Education and IZA)

May 2018

The Programme for International Student Assessment (PISA) is an important international study of 15-year-olds' knowledge and skills. New results are released every three years, and have a substantial impact upon education policy. Yet, despite its influence, the methodology underpinning PISA has received significant criticism. Much of this criticism has focused upon the psychometric scaling model used to create the proficiency scores. The aim of this paper is to therefore investigate the robustness of cross-country comparisons of PISA scores to subtle changes to the underlying scaling model used. This includes the specification of the item-response model, whether the difficulty and discrimination of items are allowed to vary across countries (item-by-country interactions) and how test questions not reached by pupils are treated. Our key finding is that these technical choices make little substantive difference to the overall country-level results.

Key Words: PISA, item-response theory.

Contact details: John Jerrim (J.Jerrim@ucl.ac.uk) Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, WC1H 0AL

1. Introduction

The Programme for International Student Assessment (PISA) is an important international study of 15-year-olds' knowledge and skills. Conducted by the Organisation for Economic Co-Operation and Development (OECD) every three years, the results are now widely anticipated by academics, journalists and public policymakers alike. Results from PISA have led to reforms of education systems across the world, including curriculum changes in Norway (Baird et al. 2011), reforms of national assessments in Japan and the Slovak Republic (Breakspear 2012), alterations to the number of teaching hours in Iceland (Wagemaker 2011) or to the complete reform of the general education act, as in Spain (*author cite*). It has consequently been described as '*the world's most important exam*' (BBC 2013), with Andreas Schleicher (the OECD director who leads the PISA study) having been described as '*the most important man in education*' by high-ranking policy officials (Gove 2013).

However, having established such an influential reputation, PISA and other international studies are coming under ever greater scrutiny. One particular line of criticism has been about how students' test scores are produced; the scaling methodology that lies behind the production of PISA's so-called 'plausible values'. Rather than simply adding up the number of correct responses students give to the test questions, the PISA study uses a complex Item-Response Theory (IRT) model to produce estimates of students' latent ability in each subject area. However, rather than producing one single ability estimate, multiple possible values are derived for each child. This series of values are known in the psychometric literature as plausible values, and capture the uncertainty we have surrounding students' latent ability. The intuition for using this complex approach is that it is impossible to thoroughly examine students in multiple different subjects (science, reading, mathematics, problem solving) within the confines of a two-hour test. Consequently, participants only take a *random* sub-sample of test questions, with the IRT model used to equate performance across different versions of the test, and plausible values designed to reflect the uncertainty in the results. Further details regarding the PISA test design are provided below.

Various authors have described how this process is opaque, with many of the potentially important technical details not fully understood outside of a narrow range of highly-specialised psychometricians (Brown and Micklewright 2004; Goldstein 2017), which may also have implications for how these data then get used (*author cite*). Others have suggested that the particular item-response model used in PISA until 2015 is overly-simplistic and does not fit the data well (Kreiner and Christensen 2014). Particular criticism has been reserved for PISA's

use of the Rasch model (Fernandez-Cano 2016), which some consider to be less sophisticated than the three-parameter item-response model used in other large-scale international assessments such as the Trends in Mathematics and Science Study (TIMSS). This has consequently led to various different opinions emerging, ranging from whether the methodology behind PISA is sufficiently transparent (Spiegelhalter 2013; Goldstein 2017) through to whether this study is actually fit for purpose (Stewart 2013).

A key question that therefore emerges from this literature is how much do the technicalities around the PISA scaling model actually matter? Not only in terms of national averages (upon which the 'international rankings' are based), but also other distributional statistics of importance, such as cross-country comparisons of high and low achievers, measures of educational inequality and the gender gap in students' performance?

Such issues have taken on particular importance since the publication of the PISA 2015 results, when a number of technical changes were made to the construction of the PISA scale scores (plausible values). This included[1]:

- *The introduction of item-by-country interactions*. A limited number of item-by-country interactions were included in the PISA scaling model for the first time. In other words, in PISA 2015 there were some country specific item parameters, allowing some items to be freely estimated by country. This meant some questions were treated as harder to answer correctly in some countries than in others (e.g some questions are now treated as 'harder' to answer correctly in England than in Scotland). The decision of where to allow item-by-country interactions was based upon item-fit statistics to determine differential item functioning[2], and thus based upon a purely statistical approach. See (OECD 2016:150-152) for further details. Such interactions were not used in PISA between 2000 and 2012[3].
- *The use of a two-parameter model*. In PISA 2015, questions were not only allowed to vary in terms of their difficulty, but also their 'discrimination' (i.e. how well each question is thought to measure students' reading/science/maths skills). This was not the case in PISA 2000 to 2012, when the discrimination parameter for each question was

[1] A further important change to PISA in 2015, not covered within this paper, is the introduction of computer-based assessment. See *author cite* for further discussion of this issue.
[2] Poorly fitting items were determined using two criteria: (a) root mean square deviation > 0.12 and (b) a mean deviation > 0.12 and < -.12.
[3] However, in these earlier cycles, some items were deleted if they did not fit the chosen IRT model sufficiently well across a large number of countries.

fixed to one (i.e. it was assumed that each reading/science/maths question measured reading/science/maths skills equally well).

- *Items that are 'not-reached' no longer contribute to the proficiency scores.* As a timed-assessment, not all students manage to reach the end of the test. In PISA 2000 to 2012, these 'not-reached' items were treated as incorrect responses when creating the scale scores[4]. This changed in PISA 2015, with the 'not-reached' items treated as missing data, and hence do not contribute to the level of each student's estimated latent ability.
- *Changes to how the item-parameters are estimated.* In PISA 2015, data from the 2006 through to the 2015 rounds were used in the calibration of the item-parameters[5]. This was different to the procedure used in previous PISA waves, when only data from the current round was used in the item-calibration process[6]. Consequently, item-parameters (e.g. item-difficulty) differ less between PISA 2015 and previous waves.

Yet, despite this collection of potentially important technical changes, little easily digestible information has been provided to consumers of the PISA data as to the likely impact they had upon cross-country comparisons. Indeed, more generally, little previous work has considered how technical changes made to the underlying scaling model affects international comparisons of students' achievements.[7] For instance, does using a two-parameter item-response model produce different cross-country comparisons than using a Rasch model? If 'not-reached' items are treated as incorrect rather than missing data, does this alter our view on which countries have the greatest levels of educational inequality (e.g. the gap between the highest and lowest achievers)? And does the inclusion of item-by-country interactions mean that cross-national differences in PISA scores become more or less pronounced? Currently, little independent information is available to consumers of the PISA results.

[4] Note that 'not-reached' items are different to 'not-answered' items. The former is where students have essentially run out of testing time and so have not seen the item. The latter refer to questions which students have seen (and thus attempted) but have not provided a response.

[5] The motivation for basing the item-parameter estimates upon the pooled 2006-2015 data was that this would maximise sample sizes at the item level, and lead to greater stability in the item-parameter estimates. As a similar approach will also be used by PISA moving forward, it should also mean that there are not sudden large changes in item parameters across different PISA cycles.

[6] A related difference is that, in PISA 2000 to 2012, only a subset of pupils in each country were used in the item-parameter calibration process. Specifically, the survey organisers randomly selected 500 students from each OECD country to form an international sub-sample, whom the item-parameter estimates were based upon.

[7] One important exception is Brown et al. (2007). Using TIMSS 1995 data, they consider how the change from a one to a three parameter item-response model impacted upon cross-country comparisons. They concluded "*cross-country patterns of central tendency to be robust to the choice of [item-response] model. But the same is not true for dispersion, for which model choice can have a big effect.*" They hence advised that "*survey reports should include an analysis of the sensitivity of basic results to model choice*" – though this suggestion has yet to be taken up.

The aim of this paper is to therefore make this important contribution to the existing evidence base. Focusing upon the results for science, the major domain in PISA 2015, we illustrate how cross-country comparisons of key distributional statistics change once specific technical aspects of the PISA scaling model are altered. This includes a consideration of all the major changes made to the scaling model in PISA 2015, as outlined above. To preview our key findings, we discover that relative differences between countries are generally unaffected by the scaling model used. This holds true not only on average, but also for key statistics frequently used to describe the distribution of students' achievement, as well as covariation with key demographic characteristics. We consequently conclude that most of the headline findings from PISA do not seem to be particularly sensitive to the scaling model used.

The paper now proceeds as follows. Section 2 describes the PISA data and the approach we have taken to replicate the official PISA science plausible values. Section 3 then reports results from this replication. In section 4 we discuss various changes we make to the PISA scaling model, with these results reported in section 5. Conclusions and directions for future research follow in section 6.

2. Data and replication of the PISA 2015 plausible values in science

The data we use are drawn from PISA 2015. Although a total of 72 countries and economies participated, we restrict our attention to the 35 members of the OECD. The focus of this paper is therefore the robustness of the PISA results within rich, developed countries. In each country, a two-stage sample design was used, with schools selected as the primary sampling unit and students then randomly selected from within. A total of around 150 schools and approximately 5,500 pupils participated within each OECD country. Response rates, after the inclusion of 'replacement schools', were around 90 percent in most countries at both the pupil and school level.

PISA employs a complex test design. In 2015, the study included 184 questions in science, 81 questions in mathematics, 103 questions in reading and 117 in collaborative problem solving. It is, of course, impossible to expect all students to provide an answer to each of these questions within the space of a two-hour test. Test questions from the different subject areas were divided into subject specific clusters, which were then organised to create around 66 different test forms. Participating students were then randomly assigned one of these forms to complete. Consequently, although all students answered one hour's worth of science questions, only around 40 percent of students answered any questions in reading, 40 percent any questions in

mathematics and 30 percent any questions in collaborative problem solving (OECD 2016:40). The survey organisers then used an extended item-response model, incorporating how students responded to each test question they were assigned plus information from the background questionnaire, to estimate a distribution of students' latent achievement in each subject area. In other words, rather than producing a single 'test score' for each child, this item-response model produces a range of possible values. 'Plausible values' are then created by the survey organisers, which are essentially random draws from each child's estimated latent achievement distribution. Further details with respect to the PISA test design can be found in OECD (2016:Chapter 2) and the item-response methodology in OECD (2016: Chapters 9 and 12).

Within this paper, we make use of the publicly available item-level data and item-parameters provided by the OECD to broadly replicate the methodology used to generate the PISA plausible values in science.[8] Specifically, we fit a multi-dimensional item response model to students' item-response data, constraining the item-parameters to the values published in the PISA 2015 technical report (OECD 2016). Following the OECD's methodology, this model allows for students' latent science, reading and mathematics abilities to be correlated, via the inclusion of correlated error terms within the measurement model. Consequently, scale scores are produced for each student in each subject area – even in those subjects where they have not answered any test questions (see *author cite* for further details about this aspect of the PISA test design).

A simplified summary of the model we estimate is presented in Figure 1. We estimate this model separately for each language group within each country, generating for each pupil their Expected A Posteriori (EAP) proficiency estimates in each subject along with their standard errors (as a measure of uncertainty). We then draw ten random values for each student from a normal distribution in order to generate our plausible values (PVs). The mean of this normal distribution is set, for each student, to their EAP achievement estimate, with the standard deviation of the distribution set to their EAP standard error. Finally, we standardise these values across the OECD, so that they have the same mean and variance as the 'official' PVs. Our focus within this paper is therefore the ***relative*** performance of countries against one another. In other words, does making a particular change to the PISA scaling procedure advantage any one country compared to another?

[8] The item-level PISA data are available from http://www.oecd.org/pisa/data/2015database/. International item parameters are available from http://www.oecd.org/pisa/data/2015-technical-report/. Information on item-by-country interactions were provided to the authors by the OECD.

Note that the OECD do not report 'official' EAP values in the international PISA database; they only include plausible values[9]. However, as plausible values contain measurement error (they are random draws), correlations between our PVs and the OECD's 'official' PVs will be attenuated. In other words, if we were to compare the correlation between our PVs and the OECD's PVs, this would underestimate how well we have managed to reproduce the PISA scaling methodology at the individual pupil level. To overcome this issue, we create proxy 'official' EAP estimates by averaging the ten 'official' PVs in the international PISA database. We then correlate our 'replicated' EAPs to these 'official' EAPs to consider how closely we have managed to replicate the OECD's scaling procedure.

<< Figure 1 >>

Although we largely follow the methodological approach of the OECD in generating the PISA plausible values, it is important that we document a handful of areas where there are some subtle differences. First, in the OECD model, all the data collected in the background questionnaire has a direct role in the generation of the PISA plausible values. Specifically, an enormous principal components analysis is conducted upon all the background variables, with the derived components then included in the model as direct effects upon students' science, reading and mathematics achievement.[10] In contrast, Figure 1 illustrates how we have only included gender as a direct background regressor in our model.[11] Second, while we have included three subjects in our multi-dimensional item response theory (IRT) model (science, reading and mathematics), the OECD version includes financial literacy and collaborative problem-solving (for those countries that participated in these national options) as well.[12] Third, whereas we have estimated separate models for each language group within all nations, the OECD did this in only a handful of countries (Belgium, Canada and Israel – see OECD 2016: Chapter 9 page 67).[13] Fourth, all of our models have been estimated using Stata (a well-known and widely used statistics package) while the 'official' scale scores were produced by

[9] EAPs and their standard errors reflect the mean and standard deviation of each child's latent proficiency distribution in a subject. PVs are, on other hand, random draws for each child's latent proficiency distribution.
[10] The principal components analysis is performed separately in each country, with the number of components retained sufficient to explain around 80 percent of the common variance in the background data. In Figure 1, these direct effects would be represented by additional squares with arrows pointing towards the circular latent achievement variables.
[11] The inclusion of additional background issues led to convergence issues in the maximum likelihood estimation in a number of countries, while in others increased estimation time to prohibitive levels.
[12] We have excluded these additional domains from our model due to (i) the data not being publicly available at the time of writing and (ii) it would require the inclusion of several additional latent correlations, increasing the complexity of the model, and hence estimation times and convergence issues.
[13] For the other countries with more than one language group, the OECD ran a single model, though this did include item-by-country interactions in the measurement model.

the Educational Testing Service (ETS) using their own specialised software ('DGROUP'). Finally, we have used maximum likelihood procedures to estimate the model underlying our replication of the PISA proficiency scores. The OECD, in contrast, used the Laplace approximation (see OECD 2016: Chapter 9).

Given these differences, how closely has our procedure replicated the 'official' PISA proficiency scores? We consider this at both the individual pupil and country levels, focusing upon the results for science (our subject of interest). Figure 2 and Table 2 provide results for the former, illustrating the correlation between our EAP science proficiency estimates and the analogous 'official' values calculated directly from the public-use PISA database.[14]

<< Table 2 >>

<< Figure 2 >>

Figure 2 illustrates how the correlation between our science EAP estimates and the average of the OECD's ten plausible values is very high (r = 0.96) when looking at students drawn from across all countries. Table 2 then extends this result to illustrate that it also holds within each individual country of interest. In other words, despite the handful of subtle differences between our scaling model and the scaling model used by the OECD, we nevertheless closely replicate students' proficiency estimates in science, as reported in the international database.

Using our replicated plausible values, are we also able to successfully reproduce the official PISA country-level results? Figure 3 provides answers this question for mean scores and other key statistics (10^{th} percentile, 90^{th} percentile and the standard deviation; see Appendix A). The correlation we find is even stronger (approximately 0.99), with the difference between our country means and those produced using the 'official' PISA plausible values typically differing by just a couple of test points. Together, the above demonstrates how we have managed to closely reproduce the official PISA science scores. Our replicated values will therefore serve as a robust baseline for us to measure change against, once we have made some technical alterations to the underlying scaling model used.

In the following section, we illustrate how cross-country comparisons change after making a number of alterations to the PISA 2015 scaling model. First, PISA 2015 allowed for a limited number of item-by-country interactions. This means that the difficulty and discrimination

[14] Note that the correlation between the average of the first five PVs and the last five PVs is 0.983. We take this as approximately the maximum possible correlation that is achievable, given the random error within the PVs.

parameters were allowed to be higher or lower in some countries than in others (usually due to concerns over poor model fit). Although the number of such interactions used in PISA 2015 was small, their inclusion in the scaling model is somewhat of a contentious issue. It has been suggested, for instance, that this may 'smooth out' important and interesting differences between countries (Goldstein 2017) and could jeopardise cross-national comparability. Likewise, on a conceptual level, it seems difficult to justify why some questions should be treated as harder in Scotland than in England (for example) – as the PISA 2015 scaling model does. We hence begin by investigating whether excluding such interactions from the PISA scaling model would lead to an appreciable change to the results.

The second change we make is to the parameterisation of the underlying IRT model. Specifically, a two-parameter model was used in PISA 2015; something that was seen as a significant departure from past waves of PISA when a Rasch model was used. Table 2 provides some descriptive information on the distribution of the discrimination parameters used in PISA 2015, illustrating how the average value was typically just over the value of one used in the Rasch model. In the following section we consider how the PISA 2015 results would look (in terms of relative differences between countries) if a Rasch model had been used instead. We return to our scaling model and constrain all the discrimination parameters to one, thereby assuming each science question measures students' science skills equally well.

<< Table 2 >>

Third, as in previous cycles of PISA, there were some non-trivial changes to the estimated item-parameters between PISA 2015 and previous cycles. Not only was the discrimination parameter allowed to vary (see Table 2), but the item-difficulty also changed. For instance, PISA 2015 used different difficulty parameters than PISA 2006 in science even for the same items (as did previous waves of PISA). But does altering the item-parameters used in the scaling model really make any difference to the results? We consider how the PISA 2015 results would change if the 2006 item-parameters were used instead (we use the parameters from 2006 as this was the only other time science was the major PISA domain). Specifically, this implies that we constrain all discrimination parameters to one (i.e. we fit a Rasch model) and use the 2006 item-difficulty values (instead of the 2015 values) where they are available. This is possible only for trend items, and not for the new science questions introduced in PISA 2015 (where we continue to use the 2015 item-parameters). The purpose of this particular

exercise is to demonstrate whether using a different set of item-parameter estimates leads to substantial changes to the cross-national pattern of results.

Fourth, the PISA 2015 scoring procedure treated 'not-reached' questions as missing data – and hence did not make any contribution to students' proficiency scores. Within our analysis we illustrate how cross-country comparisons change if these not-reached questions are treated as incorrect responses instead (as per the PISA 2000 to 2012 approach). Annex A2 provides an overview of the percentage of questions classified as 'not reached' by country and subject, illustrating that this is typically very low (less than two percent of questions being unreached). Although it is therefore unlikely that altering the treatment of not-reached items in the PISA scaling model had an impact upon average scores, it may have had an influence upon some other statistics of interest (e.g. percentage of low achievers, inequality in educational achievement). One of our primary interests will hence be how this change influences international comparisons of low-performance (e.g. the 10^{th} and 25^{th} percentiles) and measures of educational inequality (e.g. the standard deviation, socio-economic gaps), under the assumption that lower-achieving and disadvantaged students are most likely to fail to complete the test within the time limit (Bridgeman, McBride, & Monaghan, 2004).

Finally, we ask what is the cumulative impact of making all the changes outlined above? In other words, how would the relative position of countries change when multiple alterations are made to the scaling model?

To summarise the consistency of results across the different models, we use the Spearman rank correlation. This measures the direction and strength of the association between two ranked variables, and thus illustrates how the rank-ordering of countries changes when the various different alterations to the PISA scaling model are made. Country average scores and country rankings are also provided to illustrate how the alterations of the scaling model influences these particular statistics.

3. Results

Excluding item-by-country interactions

Figure 4 illustrates the correlation between our original replicated country-average science scores described in section 2 (x-axis) and our alternative estimates when the item-by-country interactions have been excluded from the scaling model (y-axis). This is complemented by the first column of Table 3, which illustrates the analogous strength of the cross-country correlations for various distributional statistics (10^{th}, 25^{th}, 50^{th}, 75 and 90^{th} percentile, mean

and standard deviation)[15]. The clear message is that whether item-by-country interactions are included or excluded from the scaling model makes essentially no difference to the substantive results. The correlation between the two sets of estimates are extremely high for all the country-level descriptive statistics considered, with all the Pearson coefficients sitting above 0.99. Hence there is no evidence that the inclusion of item-by-country interactions into the PISA scaling model has provided a particular advantage (or disadvantage) for any of the countries we consider.

<< Table 3 >>

<< Figure 4 >>

In Table 4 illustrate how this translates into changes in the mean PISA science scores across countries. Following on from the previous results, these two tables further illustrate how the removal of item-by-country interactions barely leads to any change in the results. For instance, even in countries where the movement is most extreme, the average science score changes by just three or four test points (e.g. Ireland and Switzerland). Likewise, the standard deviation varies by less than a single PISA test point in most countries if item-by-country interactions are excluded. Consequently, Table 4 helps to further illustrate how this technical feature of the PISA scaling model has almost no impact upon the substantive results.

<< Table 4 >>

Applying a 'Rasch' model

What happens to cross-country comparisons in PISA 2015 if item-discrimination is no longer allowed to vary, and a Rasch model is fitted instead? To begin, Table 5 provides a comparison of 'model fit' between our Rasch and two-parameter models, based upon the Akaike Information Criterion (AIC). The AIC is a statistic that is commonly used to decide between two competing models, and trades-off parsimony (number of estimated parameters) against how closely the model aligns with the empirical data. It is therefore a measure of *relative* fit, used to judge one model against another, with preference given to the model generating the lower AIC value[16]. Table 6 reveals that, in most countries, the AIC is lower for the two-parameter model than the Rasch model. In other words, we find evidence that the two-

[15] Note that, throughout this section, we use our replicated plausible values. (We only produced EAP estimates for the purpose of the previous section, where we investigated how well our replication worked).
[16] We have also estimated the Bayesian Information Criterion (BIC) for the two models in each country, with the same substantive conclusions reached.

parameter model introduced in PISA 2015 is typically an improvement over the Rasch model used in PISA 2000 to 2012 in terms of model fit.

<< Table 5 >>

But has this improved fit to the data led to a substantive change in the country-level results? The second column of Tables 3 and 4 provides the answer, and again illustrates how international comparisons of various descriptive statistics are largely unaffected by this choice. For instance, the mean, standard deviation and selected achievement percentiles are all virtually identical regardless of the approach used (the spearman rank correlations are all approximately 0.99). Hence, despite PISA having received a great deal of criticism for its historical use of the Rasch model, we find little evidence that moving to a more complex two-parameter item-response model has any meaningful impact upon cross-country comparisons of the results.

Using the 2006 item parameters (rather than 2015)

As well as allowing the discrimination parameter to vary, the item-difficulty parameters used in PISA 2015 also differed from previous rounds. But how much impact does using different IRT item-parameters really have upon the results? The third column of Table 3 provides insight into this issue, where we have used the 2006 values of the item-parameters in the scaling model rather than the 2015 values.[17] Consistent with the findings presented in the sub-sections above, altering the item-parameters used in the scaling model leads to only trivial changes to the estimates. In particular, note how the Spearman correlations reported are consistently very strong (approximately 0.99) for each of the distributional statistics considered. Moreover, for most countries, the average score and rank position presented in Table 4 are broadly stable. Consequently, the exact value of the item-parameters used in the scaling model (and whether a Rasch or two-parameter IRT model is used) has a trivial impact upon the substantive conclusions reached.

Treating not reached items as incorrect

In-line with the findings presented thus far, the impact of altering how 'not-reached' items are treated has a trivial impact upon cross-national comparisons of students' achievement. Importantly, this is not only true on average (mean scores) but also for comparisons of the lowest-achievers, as measured by the 10^{th} and 25^{th} percentiles of the science distribution. Specifically, the fourth column of Table 3 illustrates how the cross-country correlations

[17] Note that the use of the 2006 parameters implies that a Rasch model is fitted (i.e. we set all discrimination parameters to one).

reported are all consistently above 0.99, with almost no substantive change to countries positions in the international rankings in Table 4. We consequently conclude that this particular analytic choice has almost no impact upon the results.

The combined effect

The final column of Table 3 provides the correlations between (a) our initial replication of key country-level statistics and (b) alternative country-level estimates once all the changes made to the scaling model covered in the sub-sections above have been taken into account. Given the results presented thus far, it is perhaps unsurprising that the correlation coefficients all remain extremely high (around 0.99). Likewise, the country average science scores and rankings remain very similar between the first and last columns of Table 4. In other words, even when a raft of changes are made to the scaling model, the same cross-national pattern of results continues to be found. Consequently, this provides yet more evidence of how cross-country comparisons made within a given PISA cycle are robust to the choice of the scaling model used.

Do similar findings hold for other subject areas?

All of the estimates presented thus far relate to the results in science – the major domain in PISA 2015. Do we find similarly strong correlations for the minor domains (reading and mathematics)? Table 6 provides a summary of our results for these two subjects based upon the Spearman's rank correlation. This is supplemented by Tables 7 and 8, which illustrates how average scores and country rankings change as the various alterations to the PISA scaling model are made. Consistent with our findings for science, we find little change to the cross-country pattern of results when changes are made to the scaling model. The correlations we find remain extremely high across the various distributional statistics considered, though are slightly lower than the analogous results for science. This is likely to be due to reading and mathematics being 'minor domains' in PISA 2015, with students answering less questions on these topics, and hence the specification of scaling model having a slightly more important role. Nevertheless, the results we have presented for science throughout this section do generally seem to hold in other subject areas as well.

<< Table 7 >>

<< Table 8 >>

4. Conclusions

In this paper we have investigated whether the precise specification of the PISA scaling model really makes a substantial difference to cross-national comparisons of educational achievement. Our results provide a clear and consistent message. Even when multiple alterations are made to the scaling model, this only has a trivial impact upon cross-country comparisons within a given PISA cycle. This holds true across a range of key statistics (mean, standard deviation, gender differences) and the different PISA domains (science, reading and mathematics).

There are two potential ways of interpreting these findings. First, there is a view within parts of the psychometric community that the scaling model used in previous rounds of PISA was flawed, particularly with respect to the use of the Rasch model (Kreiner and Christensen 2014). Yet, given that we have shown cross-country comparisons do not really change when a more complex methodology is used, it was perhaps good enough, and that some of the media reports questioning this aspect of the study have been overblown. Alternatively, one might conclude that the new methodology introduced in PISA 2015 is therefore equally as flawed as the methodology used before, given that it does not produce substantially different results. Our own view is closer to the former – we believe our investigations illustrate how the key results from PISA (at least as far as the psychometric scaling model are concerned) seem to be relatively robust to the technical choices made. Nevertheless, we believe further investigations in the spirit of those conducted within this paper should be welcomed by the OECD and the scientific community to further justify the chosen psychometric approach.

These findings should, of course, be interpreted in light of the limitations of this paper, and the need for further research. First, this paper has focused solely upon relative differences between OECD countries within a single PISA cycle. We have not considered how the scaling approach influences absolute measures of students' performance, such as changes in a country's PISA scores over time, or for middle and low-income countries. Although clearly a topic of great importance, it is beyond the scope of this paper, but remains an important area for future research. Second, we have focused upon a particular set of changes made to the scaling model, motivated by the fact that these technical details have altered across the PISA cycles (most notably in 2015). Although these changes are quite extensive from a psychometric perspective, including much debated issues in this technical literature (e.g. the impact of shifting from a Rasch to a two-parameter model), we can obviously not rule out the possibility that making

some other changes may have some kind of an impact upon the results (e.g. if PISA were to move to a three-parameter IRT model instead).

Despite these limitations, we believe this paper has made an important contribution to ongoing debates about PISA and other large-scale assessments. Although there are clearly important limitations to such studies, our analysis suggests that some of the criticisms made of the scaling methodology are unjustified. Although a complex methodology is used, one which is not widely understood outside a highly-specialised psychometric field, the scaling model can be closely replicated using information freely available in the public domain. More importantly, cross-country comparisons seem to be largely unaffected by the precise specification of the scaling model used. By completing this independent investigation, it is hoped that this will be accurately reflected in media reports of future PISA results, and that there is a greater appreciation amongst sceptics that international comparisons seem quite robust to departures from the official OECD scaling approach.

References

Baird, J.; Isaacs, T.; Johnson, S.; Stobart, G.; Yu, G.; Sprague, T. and Daugherty, R. 2011. 'Policy effects of PISA.' Oxford University Centre for Educational Assessment working paper. Accessed 10/02/2017 from http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Breakspear, S. 2012. 'The policy impacts of PISA'. *OECD working paper no. 71*. OECD: Paris.

Bridgeman, B., McBride, A. & Monaghan, W. 2004. Testing and time limits. R&D Connnections, Educational Testing Services. Accessed 26/03/2018 from *https://www.ets.org/Media/Research/pdf/RD_Connections1.pdf.*

Brown, G. and Micklewright J. 2004. 'Using international surveys of achievement and literacy: a view from the outside.' UNESCO working paper. Accessed 09/02/2017 from http://www.uis.unesco.org/Library/Documents/wp2-04-en.pdf

Brown, G, Micklewright, J, Schnepf, S.V., and Waldmann, R. 2007. 'Cross-national comparisons of learning achievement: How robust are the findings?' *Journal of the Royal Statistical Society Series A* 170(3): 623-646.

Fernandez-Cano, A. 2016. 'A methodological critique of the PISA evaluations.' *RELIEVE 22(2): M15*

Goldstein, H. 2017. 'Measurement and evaluation issues with PISA.' In L. Volante (Ed) *The PISA effect on global educational governance* (pp 49-58). Routledge.

Gove, M. 2013. 'The Secretary of State for Education sets out the importance of comparing English schools with the best in the world.' Accessed 26/11/2017 https://www.gov.uk/government/speeches/michael-gove-article-in-the-independent-on-education-reforms

Kreiner, S. and Christensen, K. 2014. 'Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy.' *Psychometrika* 79(2): 210-231.

OECD. 2016. 'PISA 2015 technical report.' Accessed 28/03/2018 from http://www.oecd.org/pisa/data/2015-technical-report/

Oliveri, M and von Davier, M. 2011. 'Investigation of model fit and score scale comparability in international assessments.' *Psychological Test and Assessment Modeling* 53 (3): 315-333.

Spiegelhalter, D. 2013. 'The problems with PISA statistical methods.' *StatsLife. Accessed 26/11/2017 from https://www.statslife.org.uk/features/1074-the-problems-with-pisa-statistical-methods*

Stewart, W. 2013. 'PISA rankings are utterly wrong.' *Times Educational Supplement*. Accessed 26/11/2017 from https://www.tes.com/news/tes-archive/tes-publication/pisa-rankings-are-utterly-wrong#.Ue49u5PAJzk.twitter

Stewart, W. 2013b. 'Is PISA fundamentally flawed?' *Times Educational Supplement*. Accessed 26/11/2017 from https://www.tes.com/news/tes-archive/tes-publication/pisa-fundamentally-flawed

Wagemaker, H. 2011. IEA: International studies, impact and transition. In C. Papanastasiou, T. Plomp and E. Papanstasiou (Eds.) *IEA 1958-2008: 50 years of experiences and memories* (pp 253-273). Nicosia:Cultural Center of the Kykkos Monatery.

Table 1. Correlation between our EAP science scores and the OECD EAP science scores at the pupil level. Results by country.

Country	Correlation
Switzerland	0.965
New Zealand	0.964
Australia	0.964
Norway	0.963
Luxemburg	0.961
Sweden	0.961
France	0.960
Canada	0.960
Austria	0.959
Israel	0.958
Ireland	0.958
United Kingdom	0.957
Japan	0.956
Poland	0.954
Hungary	0.954
Finland	0.953
Czech Republic	0.953
USA	0.952
Greece	0.951
Netherlands	0.951
South Korea	0.951
Germany	0.951
Estonia	0.950
Belgium	0.950
Portugal	0.947
Israel	0.947
Hong Kong	0.946
Italy	0.945
Slovenia	0.943
Spain	0.943
Slovak Republic	0.943
Denmark	0.939
Latvia	0.938
Turkey	0.932
Chile	0.929
Mexico	0.894
Average	**0.950**

Notes: The OECD EAP estimate is approximated as the average of the ten plausible values in science for each student.

Table 2. The distribution of the discrimination item parameters used in PISA 2015

	Science	Mathematics	Reading
10th percentile	0.68	0.74	0.70
25th percentile	0.90	1.00	0.83
50th percentile	1.08	1.00	1.00
75th percentile	1.42	1.27	1.15
90th percentile	1.65	1.70	1.44
Mean	**1.15**	**1.12**	**1.02**
Standard deviation	**0.39**	**0.38**	**0.28**

Notes: Authors calculations using data from the PISA 2015 technical report.

Table 3. A summary of the correlation between country estimates of selected distributional statistics for science after making changes to the scaling model used

	1. No unique parameters	2. Rasch model	3. 2006 item parameters	4. Not reached as incorrect	5. All changes
Mean	0.995	0.995	0.938	0.994	0.993
St dev	0.988	0.982	0.984	0.993	0.983
P10	0.990	0.992	0.941	0.990	0.989
P25	0.995	0.991	0.938	0.992	0.995
P50	0.995	0.995	0.940	0.995	0.995
P75	0.985	0.986	0.933	0.983	0.980
P90	0.992	0.984	0.945	0.988	0.983
Gender gap	0.997	0.994	0.954	0.996	0.994
SES gap	1.000	0.999	0.998	1.000	0.999

Notes: Figures refer to the correlations at the country level. All comparisons refer to correlations with our original replicated values of the PISA science scores. These have been calculated using our replicated plausible values.

Table 4. How do mean science scores change when the PISA scaling model is altered?

	0. Replicated values	1. No unique parameters	2. Rasch model	3. 2006 item parameters	4. Not reached as incorrect	5. All changes
Japan	537.8 (1)	538.8 (1)	537.6 (1)	539.3 (1)	538.2 (1)	538.6 (1)
Estonia	532.9 (2)	533.7 (2)	534.9 (2)	533.8 (2)	533.4 (2)	533.4 (2)
Finland	530.7 (3)	530.5 (3)	533.8 (3)	532.7 (3)	531.4 (3)	532.7 (3)
Canada	528.9 (4)	525.6 (4)	525.7 (4)	525.2 (4)	525.3 (4)	524.8 (4)
Germany	513.9 (5)	510.6 (8)	512.1 (6)	510.8 (8)	509.2 (8)	509.4 (9)
New Zealand	513.7 (6)	510.9 (7)	511.1 (8)	511.3 (7)	509.3 (7)	510.1 (7)
Slovenia	513.3 (7)	513.2 (6)	516.1 (5)	514.6 (5)	513.7 (5)	514.1 (5)
South Korea	512.5 (8)	513.9 (5)	511.3 (7)	512.2 (6)	512.6 (6)	511.9 (6)
Switzerland	511.7 (9)	508.0 (10)	509.0 (10)	508.1 (10)	507.1 (10)	506.9 (10)
Netherlands	511.3 (10)	508.5 (9)	509.5 (9)	509.7 (9)	507.8 (9)	510.0 (8)
Australia	509.1 (11)	507.0 (11)	505.9 (11)	506.8 (11)	505.7 (11)	506.6 (11)
Ireland	508.6 (12)	504.7 (13)	503.9 (14)	504.0 (15)	505.7 (12)	504.0 (13)
Belgium	507.8 (13)	504.8 (12)	503.3 (15)	504.3 (13)	503.4 (13)	503.4 (15)
Denmark	506.2 (14)	503.9 (14)	504.3 (13)	505.8 (12)	501.8 (15)	504.8 (12)
Norway	503.7 (15)	500.0 (17)	502.2 (16)	500.7 (17)	499.0 (17)	499.1 (17)
Poland	502.5 (16)	502.5 (15)	505.6 (12)	504.1 (14)	502.8 (14)	503.5 (14)
Austria	501.9 (17)	498.2 (18)	500.9 (18)	499.0 (18)	498.2 (18)	498.2 (18)
Portugal	500.6 (18)	501.5 (16)	501.1 (17)	500.8 (16)	500.7 (16)	500.1 (16)
Spain	497.2 (19)	496.0 (19)	497.3 (19)	495.3 (20)	497.5 (19)	494.8 (21)
France	495.7 (20)	495.9 (20)	495.5 (21)	495.3 (21)	493.9 (22)	493.1 (22)
United Kingdom	494.9 (21)	495.4 (21)	494.2 (23)	494.3 (23)	494.6 (21)	493.1 (23)
USA	493.9 (22)	495.0 (22)	495.0 (22)	494.7 (22)	495.5 (20)	495.7 (20)
Czech Republic	492.6 (23)	494.3 (23)	495.9 (20)	496.5 (19)	492.9 (23)	495.9 (19)
Sweden	492.3 (24)	493.6 (24)	492.4 (24)	493.4 (24)	490.7 (24)	491.1 (24)
Latvia	488.4 (25)	488.8 (25)	487.6 (25)	488.4 (25)	489.7 (25)	489.1 (25)
Italy	484.6 (26)	485.0 (26)	486.5 (26)	-	483.6 (26)	483.3 (26)
Luxemburg	482.4 (27)	483.3 (27)	484.2 (28)	483.5 (27)	482.2 (27)	482.6 (27)
Hungary	480.2 (28)	479.8 (28)	484.8 (27)	482.6 (28)	480.8 (28)	482.3 (28)
Iceland	475.3 (29)	475.2 (29)	475.9 (29)	474.8 (29)	475.0 (29)	473.8 (29)
Israel	469.8 (30)	469.5 (30)	468.7 (30)	468.9 (30)	468.8 (30)	467.2 (30)
Slovak Republic	463.2 (31)	463.0 (31)	464.4 (31)	464.6 (31)	463.9 (31)	464.7 (31)
Greece	456.5 (32)	454.4 (32)	453.6 (32)	454.0 (32)	453.6 (32)	453.4 (32)
Chile	447.6 (33)	450.0 (33)	450.6 (33)	450.5 (33)	448.7 (33)	448.4 (33)
Turkey	426.7 (34)	426.1 (34)	425.1 (34)	426.8 (34)	428.0 (34)	427.6 (34)
Mexico	415.1 (35)	415.8 (35)	416.1 (35)	415.8 (35)	416.8 (35)	416.9 (35)

Notes: The model using 2006 item parameters for Italy did not converge to a solution. Italy and the 26th rank is therefore excluded from these results.

Table 5. A comparison of the Rasch and two-parameter model fit across countries

	AIC		
	Rasch	**2PL**	**Change**
Australia	21,265,466	21,242,162	-23,304
Austria	6,164,697	6,156,881	-7,817
Belgium (Flemish)	5,130,234	5,099,247	-30,987
Belgium (French)	4,240,548	4,202,457	-38,091
Canada (English)	22,216,850	22,185,880	-30,970
Canada (French)	5,901,455	5,888,759	-12,696
Chile	15,348,660	15,125,470	-223,190
Czech Republic	7,055,822	7,049,863	-5,959
Denmark	5,133,536	5,133,155	-381
Estonia	722,082	720,668	-1,414
Finland	4,804,141	4,801,951	-2,190
France	59,940,784	59,876,552	-64,232
Germany	62,169,884	62,075,812	-94,072
Greece	7,947,821	7,946,050	-1,772
Hungary	6,925,020	6,909,926	-15,094
Iceland	333,850	333,996	146
Ireland	5,871,638	5,856,239	-15,399
Israel (Hebrew)	7,123,462	7,111,920	-11,542
Israel (Arabic)	2,089,818	2,098,069	8,251
Italy	40,787,848	40,737,780	-50,068
Japan	92,960,728	92,674,440	-286,288
Latvia	998,112	997,057	-1,055
Luxemburg (German)	332,516	331,926	-590
Luxemburg (French)	119,330	119,344	14
Mexico	99,356,784	99,636,376	279,592
Netherlands	16,101,700	16,066,891	-34,809
New Zealand	4,439,132	4,429,347	-9,785
Norway	4,815,245	4,798,950	-16,295
Poland	33,830,036	33,778,144	-51,892
Portugal	8,015,001	8,009,994	-5,007
Slovak Republic	4,031,185	4,032,822	1,638
Slovenia	1,418,376	1,416,976	-1,400
South Korea	48,950,056	48,876,152	-73,904
Spain	33,712,308	33,658,860	-53,448
Sweden	7,417,497	7,405,656	-11,841
Switzerland (German)	5,448,933	5,121,406	-327,527
Switzerland (Italian)	7,357,683	318,939	-7,038,744
Switzerland (French)	5,378,396	2,257,293	-3,121,103
Turkey	481,907	484,049	2,142
United Kingdom	1,163,056	1,161,800	-1,256
USA	477,029	476,570	-458

Notes: Figures refer to the Akaike Information Criterion (AIC). Highlighted cell illustrates the lower value (and hence the preferred model).

Table 6. A summary of the correlation between country estimates of selected distributional statistics for mathematics and reading after making changes to the scaling model used

(a) Mathematics

	1. No unique parameters	2. Rasch model	3. 2006 item parameters	4. Not reached as incorrect	5. All changes
Mean	0.99	0.99	0.93	0.99	0.98
St dev	0.95	0.92	0.92	0.93	0.88
P10	0.99	0.99	0.91	0.99	0.99
P25	0.99	0.99	0.92	0.99	0.99
P50	0.98	0.99	0.93	0.99	0.98
P75	0.97	0.98	0.95	0.99	0.97
P90	0.98	0.98	0.96	0.98	0.96
Gender gap	0.92	0.93	0.93	0.89	0.86
SES gap	0.99	0.98	0.98	0.99	0.98

(b) Reading

	1. No unique parameters	2. Rasch model	3. 2006 item parameters	4. Not reached as incorrect	5. All changes
Mean	0.98	0.99	0.94	0.99	0.97
St dev	0.96	0.97	0.93	0.98	0.93
P10	0.98	0.99	0.96	0.98	0.96
P25	0.98	0.99	0.95	0.99	0.97
P50	0.98	0.99	0.94	0.99	0.97
P75	0.99	0.99	0.94	0.99	0.99
P90	0.98	0.97	0.93	1.00	0.98
Gender gap	0.91	0.94	0.90	0.93	0.90
SES gap	0.99	0.99	0.99	0.98	0.98

Notes: Figures refer to the correlations at the country level. All comparisons refer to correlations with our original replicated values of the PISA reading and mathematics scores.

Table 7. How do mean mathematics scores change when the PISA scaling model is altered?

	0. Replicated values	1. No unique parameters	2. Rasch model	3. 2006 item parameters	4. Not reached as incorrect	5. All changes
Japan	523.8 (1)	525.6 (1)	525.0 (1)	525.2 (1)	523.6 (1)	523.2 (1)
Estonia	511.7 (2)	513.5 (2)	513.8 (2)	512.8 (2)	512.4 (4)	511.9 (2)
South Korea	510.4 (3)	513.0 (3)	510.9 (4)	509.2 (4)	518.0 (2)	509.9 (4)
Canada	509.8 (4)	511.2 (4)	511.3 (3)	510.3 (3)	515.3 (3)	510.5 (3)
Switzerland	505.1 (5)	510.1 (5)	508.5 (5)	508.2 (5)	506.2 (5)	508.1 (5)
Finland	502.0 (6)	504.3 (6)	505.2 (8)	505.0 (6)	502.4 (6)	503.9 (7)
Slovenia	501.6 (7)	503.0 (9)	504.3 (9)	503.4 (7)	501.8 (8)	501.9 (8)
Poland	500.8 (8)	496.9 (14)	498.3 (13)	496.9 (14)	501.4 (9)	495.6 (14)
Denmark	500.6 (9)	503.4 (8)	508.0 (6)	502.0 (8)	502.0 (7)	500.5 (10)
Belgium	497.8 (10)	503.6 (7)	502.3 (10)	499.9 (12)	498.1 (12)	500.6 (9)
Germany	497.5 (11)	501.7 (10)	501.0 (11)	500.6 (10)	498.5 (11)	505.0 (6)
Netherlands	497.4 (12)	501.1 (11)	505.5 (7)	501.1 (9)	499.0 (10)	500.5 (11)
Ireland	494.6 (13)	497.7 (12)	499.0 (12)	496.6 (15)	497.1 (13)	496.0 (12)
New Zealand	494.3 (14)	497.1 (13)	498.2 (14)	497.3 (13)	496.3 (14)	495.9 (13)
United Kingdom	492.1 (15)	493.7 (15)	494.8 (15)	492.7 (18)	492.3 (16)	491.7 (16)
France	490.9 (16)	492.0 (20)	492.4 (18)	493.5 (16)	488.7 (21)	489.6 (21)
Australia	490.5 (17)	493.4 (16)	493.9 (16)	500.1 (11)	492.6 (15)	492.3 (15)
Portugal	490.2 (18)	491.3 (21)	492.5 (17)	492.7 (19)	489.5 (19)	490.5 (18)
Norway	488.3 (19)	492.2 (19)	491.3 (19)	493.3 (17)	489.1 (20)	490.5 (19)
Austria	488.0 (20)	492.5 (18)	491.3 (20)	492.2 (20)	489.7 (18)	490.4 (20)
Czech Republic	486.5 (21)	488.6 (22)	487.8 (21)	488.0 (21)	486.4 (23)	486.3 (22)
Spain	485.5 (22)	486.7 (23)	486.3 (23)	486.2 (24)	489.8 (17)	484.5 (23)
Italy	484.5 (23)	485.5 (24)	486.7 (22)	-	483.1 (24)	482.2 (24)
Sweden	484.2 (24)	493.0 (17)	484.5 (24)	487.7 (22)	487.9 (22)	491.3 (17)
Luxemburg	479.9 (25)	482.2 (25)	481.1 (25)	480.9 (25)	479.2 (25)	480.2 (25)
USA	477.8 (26)	474.1 (29)	473.5 (29)	474.2 (29)	474.5 (28)	474.6 (29)
Iceland	477.1 (27)	479.6 (26)	478.1 (26)	479.5 (26)	476.5 (27)	477.6 (26)
Latvia	473.7 (28)	477.6 (27)	474.7 (28)	479.0 (27)	476.5 (26)	476.8 (27)
Hungary	472.7 (29)	474.9 (28)	475.3 (27)	474.3 (28)	473.3 (29)	475.4 (28)
Slovak Republic	469.3 (30)	468.1 (31)	466.6 (30)	467.8 (30)	466.7 (30)	467.5 (30)
Israel	463.8 (31)	470.8 (30)	464.4 (31)	466.0 (31)	462.6 (31)	464.1 (31)
Greece	454.7 (32)	453.5 (32)	453.7 (32)	454.0 (32)	452.3 (32)	452.7 (32)
Chile	433.5 (33)	434.3 (33)	439.6 (33)	436.8 (33)	433.7 (33)	440.3 (33)
Turkey	424.6 (34)	421.4 (34)	423.4 (34)	423.2 (34)	425.9 (34)	423.5 (34)
Mexico	414.5 (35)	415.8 (35)	416.0 (35)	418.8 (35)	416.6 (35)	420.1 (35)

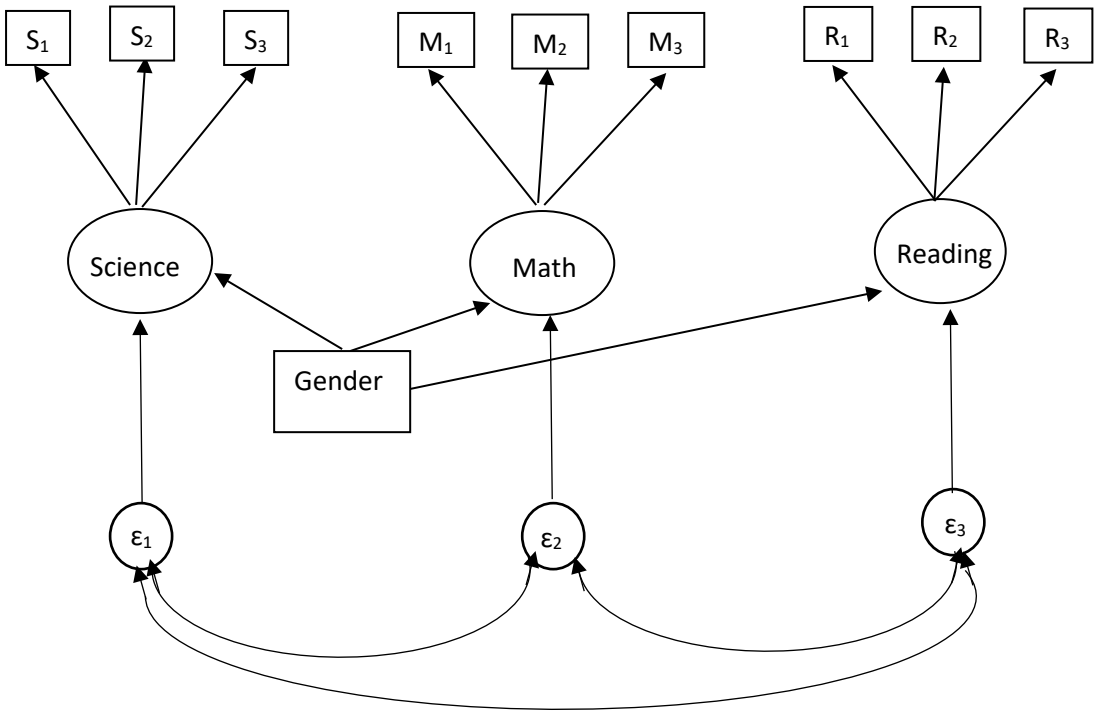
Notes: The model using 2006 item parameters for Italy did not converge to a solution. Italy and the 23rd rank is therefore excluded from these results. Figures in brackets refers to country ranking.

Table 8. How do mean reading scores change when the PISA scaling model is altered?

	0. Replicated values	1. No unique parameters	2. Rasch model	3. 2006 item parameters	4. Not reached as incorrect	5. All changes
Japan	519.8 (1)	517.7 (2)	520.0 (1)	515.8 (2)	518.0 (1)	515.8 (2)
Finland	519.0 (2)	519.9 (1)	518.7 (2)	517.2 (1)	517.7 (2)	517.5 (1)
Canada	518.9 (3)	517.5 (3)	515.6 (3)	514.8 (3)	513.0 (6)	514.8 (3)
Ireland	518.2 (4)	514.5 (5)	512.1 (6)	513.4 (4)	516.4 (3)	514.0 (4)
South Korea	516.0 (5)	510.9 (6)	514.1 (5)	508.8 (6)	514.4 (4)	509.5 (6)
Estonia	514.8 (6)	515.6 (4)	514.2 (4)	513.3 (5)	513.9 (5)	513.8 (5)
Germany	507.9 (7)	506.8 (7)	507.1 (7)	504.6 (7)	505.0 (8)	501.0 (9)
Netherlands	507.4 (8)	504.7 (8)	505.4 (8)	502.2 (8)	505.7 (7)	503.2 (7)
Norway	506.2 (9)	503.7 (9)	500.8 (12)	501.1 (9)	502.5 (9)	500.2 (11)
New Zealand	505.3 (10)	503.1 (10)	501.4 (10)	500.8 (11)	502.5 (10)	500.8 (10)
Slovenia	501.2 (11)	500.8 (12)	501.5 (9)	499.7 (12)	500.1 (11)	500.1 (12)
United Kingdom	500.6 (12)	499.8 (13)	498.8 (14)	497.5 (14)	499.3 (12)	497.8 (15)
Switzerland	500.6 (13)	498.9 (15)	498.6 (15)	497.5 (15)	498.0 (13)	496.9 (16)
Belgium	499.8 (14)	498.3 (16)	499.1 (13)	495.9 (17)	497.2 (15)	495.2 (18)
Denmark	499.4 (15)	499.0 (14)	500.8 (11)	498.2 (13)	496.5 (16)	498.0 (13)
Portugal	498.2 (16)	497.0 (18)	498.5 (16)	497.0 (16)	497.5 (14)	497.8 (14)
Australia	498.2 (17)	498.0 (17)	495.6 (18)	490.8 (22)	495.9 (17)	495.7 (17)
Poland	496.8 (18)	501.9 (11)	498.4 (17)	500.9 (10)	495.8 (18)	501.6 (8)
France	494.6 (19)	493.9 (20)	492.4 (21)	491.6 (21)	490.9 (20)	489.8 (21)
Spain	494.5 (20)	493.9 (19)	492.9 (20)	493.3 (19)	488.9 (21)	493.6 (20)
Sweden	494.0 (21)	490.9 (22)	493.1 (19)	491.8 (20)	487.1 (25)	486.2 (25)
Austria	489.3 (22)	487.0 (25)	488.0 (25)	486.0 (25)	487.1 (24)	486.4 (24)
Italy	489.2 (23)	488.8 (24)	489.2 (24)	-	487.8 (23)	486.8 (23)
Czech Republic	489.0 (24)	489.1 (23)	489.6 (23)	488.9 (24)	488.1 (22)	489.5 (22)
USA	489.0 (25)	493.8 (21)	492.1 (22)	493.4 (19)	494.1 (19)	493.9 (19)
Latvia	487.8 (26)	486.1 (26)	483.8 (26)	482.9 (26)	485.9 (26)	485.8 (26)
Luxemburg	482.1 (27)	481.7 (27)	480.5 (27)	480.9 (27)	480.9 (27)	480.5 (27)
Israel	478.6 (28)	474.9 (28)	476.7 (29)	477.3 (28)	476.5 (28)	475.6 (29)
Hungary	476.1 (29)	473.8 (30)	477.8 (28)	476.3 (29)	475.5 (29)	475.9 (28)
Iceland	473.9 (30)	474.6 (29)	471.4 (30)	473.3 (30)	472.4 (30)	473.1 (30)
Greece	463.8 (31)	467.3 (31)	466.7 (31)	468.1 (31)	467.3 (31)	468.2 (31)
Chile	462.8 (32)	462.4 (32)	459.6 (32)	465.2 (32)	460.7 (32)	458.7 (32)
Slovak Republic	456.2 (33)	456.6 (33)	457.1 (33)	457.7 (33)	457.8 (33)	458.4 (33)
Turkey	444.8 (34)	442.3 (34)	444.4 (34)	445.2 (34)	444.2 (34)	446.1 (34)
Mexico	434.5 (35)	430.2 (35)	432.4 (35)	438.0 (35)	434.2 (35)	439.0 (35)

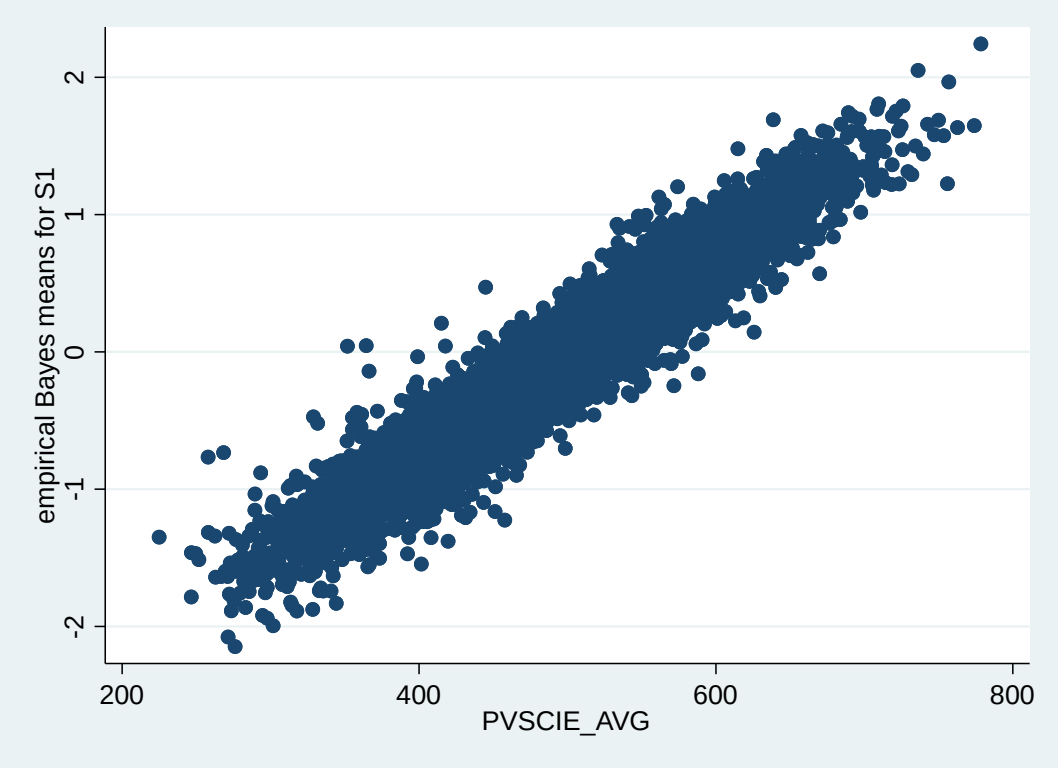
Notes: The model using 2006 item parameters for Italy did not converge to a solution. Italy and the 23rd rank is therefore excluded from these results. Figures in brackets refers to country ranking.

Figure 1. A simplified illustration of the model used to replicate the official PISA proficiency estimates (plausible values)



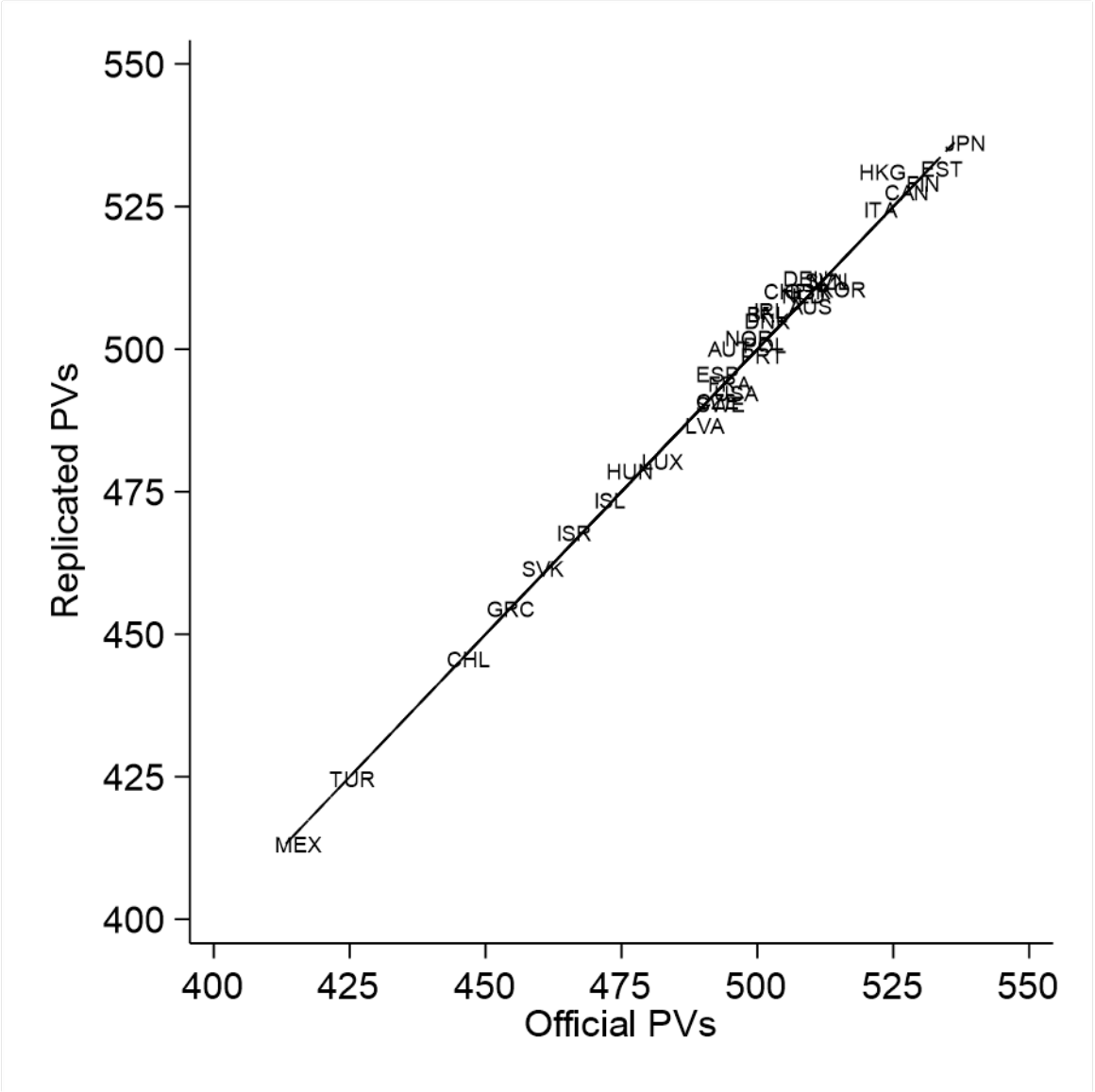
Notes: Squares refer to observed variables, ovals to latent variables and circles to error terms. S_i , M_i and R_i refer to students' responses to the PISA test questions, where 'i' is the number of items on PISA (only three items in each domain have been included in the diagram for ease of presentation). Curved lines connecting errors indicate correlated errors.

Figure 2. Correlation between our EAP science scores and the OECD EAP science scores at the pupil level



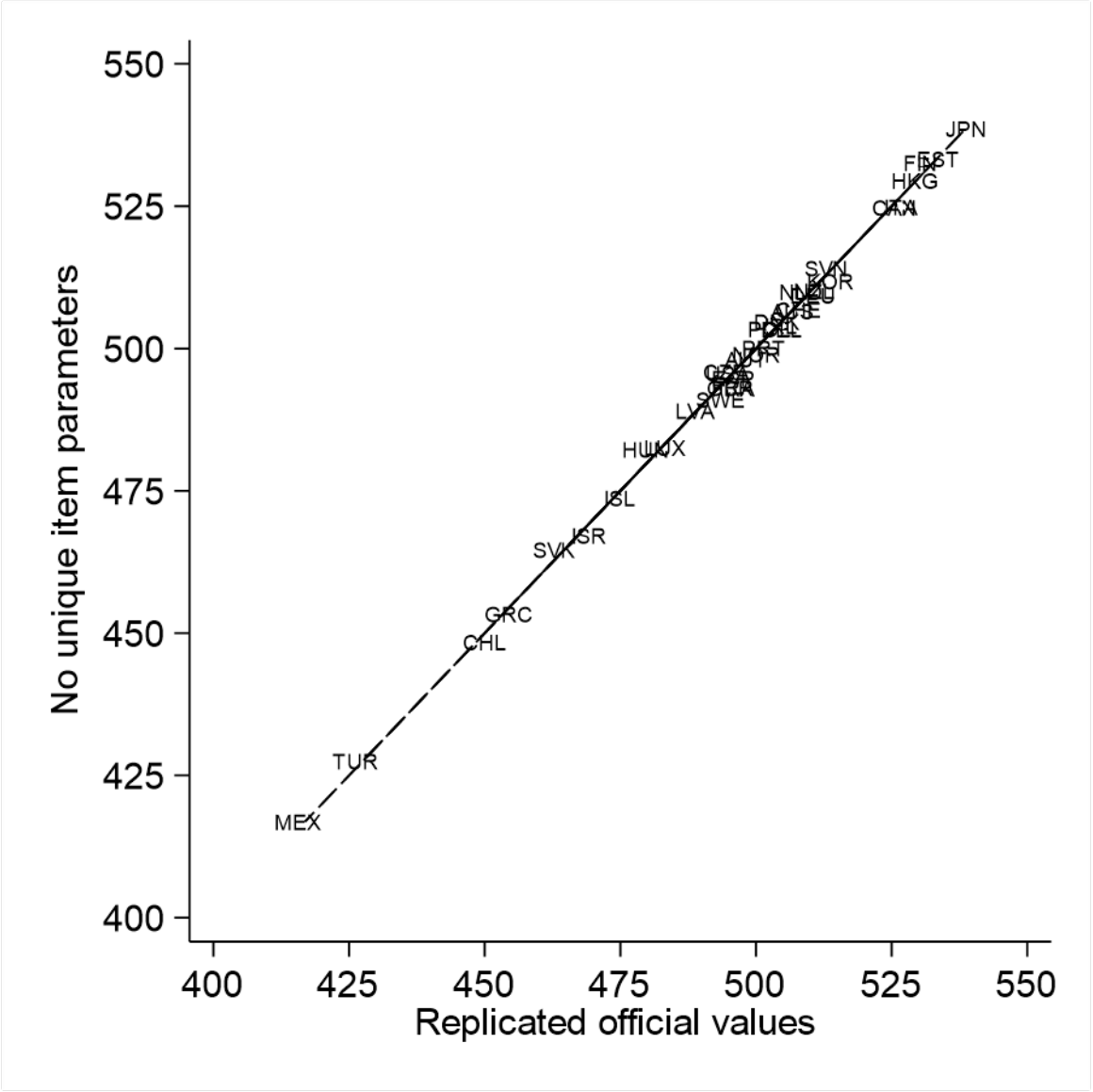
Notes: The OECD EAP estimate is approximated by taking the average of the ten plausible values in science for each student. The Pearson correlation is 0.9557 and Spearman correlation is 0.9585. Graph presented based upon a random sample of 5,000 students from the countries analysed.

Figure 3. Correlation between our estimate of the *mean* EAP science score and the OECD *mean* EAP science score at the country level



Notes: The Pearson correlation is 0.994 and Spearman correlation is 0.986.

Figure 4. A comparison of mean science score with and without including item-by-country interactions



Notes: Figure can be cross-referenced with the statistic in the top-left hand corner of Table 3 (“1. No unique parameters”). The Pearson cross-country correlation is 0.999.

Annex A1. The number of questions with unique item-parameters by country and language group

Country	Language	Math	Reading	Science
Australia	English	2	9	9
Austria	Austrian	7	5	9
Belgium (Flemish)	Flemish	1	11	13
Belgium (French)	French	2	8	9
Canada	English	3	8	9
Switzerland	German	6	11	15
Switzerland	French	2	4	9
Chile	Spanish	4	7	19
Czech Republic	Czech	1	8	15
Germany	German	3	15	8
Denmark	Danish	3	8	17
Spain	Spanish	2	5	9
Spain	Catalan	0	1	0
Spain	Valencian	0	1	0
Estonia	Estonian	3	14	21
Estonia	Russian	0	0	1
Finland	Finnish	4	14	25
France	French	3	9	12
Greece	Greek	1	14	18
Croatia	Croatian	3	8	20
Hungary	Hungarian	3	10	18
Ireland	English	3	12	26
Iceland	Icelandic	3	8	19
Israel	Hebrew	6	10	28
Israel	Arabic	0	0	8
Italy	Italian	3	13	13
Japan	Japanese	17	19	38
Korea	Korean	11	20	44
Luxemburg	German/English	3	4	4
Luxemburg	French	0	0	4
Latvia	Latvian	4	12	20
Latvia	Russian	0	0	14
Mexico	Spanish	2	8	19
Netherlands	Dutch	4	18	16
Norway	Bokmål	3	12	13
New Zealand	English	4	10	13
Poland	Polish	3	11	17
Portugal	Portuguese	2	7	14
Scotland	English	3	11	18
England/Wales/NI	English	0	13	30
Slovakia	Slovak	5	11	18
Slovenia	Slovenian	2	10	26
Sweden	Swedish	5	6	16
Turkey	Turkish	7	13	30
USA	English	3	10	13
Total # of items in PISA test		**81**	**103**	**184**

Annex A2. The percentage of 'not-reached' items by country

Country	Science %	Mathematics %	Reading %
Australia	1.1	1.2	0.7
Austria	1.5	2.0	0.9
Belgium	1.5	1.4	0.8
Canada	0.9	0.8	0.6
Switzerland	1.6	1.5	1.0
Chile	2.6	2.1	2.0
Czech Republic	1.4	2.1	0.8
Germany	1.5	1.9	0.9
Denmark	1.3	1.4	0.8
England	1.3	1.2	0.7
Spain	1.3	1.5	0.8
Finland	1.0	1.6	0.8
France	2.3	2.4	1.8
United Kingdom	1.2	1.2	0.6
Greece	1.6	1.8	1.2
Hungary	2.0	1.8	0.8
Ireland	1.6	0.8	0.5
Iceland	1.7	1.6	1.0
Israel	2.1	2.1	1.2
Italy	2.0	1.9	0.9
Japan	1.2	1.1	0.9
Korea	1.2	1.4	0.5
Luxemburg	1.7	1.5	0.9
Latvia	1.5	1.3	0.6
Mexico	1.3	1.1	1.2
Northern Ireland	1.9	1.6	1.1
Netherlands	0.7	1.0	0.4
Norway	1.6	2.0	1.4
New Zealand	1.7	1.8	1.2
Poland	1.7	1.3	0.5
Portugal	1.4	1.5	0.5
Scotland	1.8	2.7	1.1
Slovakia	1.6	1.3	0.6
Slovenia	1.1	1.3	0.5
Sweden	2.2	2.9	1.5
Turkey	1.5	1.4	0.7
USA	0.8	0.8	0.6
Wales	2.3	2.3	1.5
Average	**1.6**	**1.6**	**0.9**