Conditioning: How background variables can influence PISA scores

Laura Raffaella Zieger[a]*, J. Jerrim[a], J. Anders[b] and N. Shure[a]

[a]*Social Research Institute, University College London, London, UK;* [b]*Centre for Education Policy and Equalising Opportunities, University College London, London, UK*

*l.zieger@ucl.ac.uk; Social Research Institute, Quantitative Social Science, 55-59 Gordon Square, London WC1H 0NU

Conditioning: How background variables can influence PISA scores

The OECD's Programme for International Student Assessment (PISA) has become one of the key studies for evidence-based education policymaking across the globe. PISA has however received a lot of methodological criticism, including how the test scores are created. The aim of this paper is to investigate the so-called 'conditioning model', where background variables are used to derive student achievement scores, and the impact it has upon the PISA results. This includes varying the background variables used within the conditioning model and analysing its impact upon countries relatively positions in the PISA rankings. Our key finding is that the exact specification of the conditioning model matters; cross-country comparisons of PISA scores can change quite dramatically depending upon the statistical methodology used.

Keywords. Educational Assessment, PISA, Item Response Theory

Introduction

The Programme for International Student Assessment (PISA) is an important international study that compares mathematics, science and reading skills of 15-year-olds across countries. It has been conducted every three years since 2000 and has become the largest and most influential study of educational achievement across the world. After the publication of the PISA results, national and international stakeholders study the scores to determine who the 'winners' and 'losers' are (Sellar & Lingard, 2013). The results from PISA have consequently led to governments across the world making substantial changes to their education system. For instance, after the 'PISA shock' in Germany in 2000, major changes were made to school curricula (Ertl, 2006). Many other countries, such as Japan (Takayama, 2008), Denmark (Egelund, 2008) and other European countries (Grek, 2009), have undertaken similar reforms based upon their PISA results. PISA has hence become a source of soft educational governance, with policymakers across the world keeping a close eye upon the results.

Yet despite the impact PISA has had over the last two decades, it has not been without its critics. While some ethical concerns about the administration of PISA have been raised (e.g. Meyer, 2014), it is the methodology underpinning the study that has perhaps sparked most controversy. As discussed by Rutkowski and Rutkowski (2016) and others (Gillis et al., 2016; Hopmann et al., 2007) this includes issues such as sample representativeness, non-response rates, population coverage and cross-cultural comparability. For instance, in the case of Portugal, Freitas et al. (2016) found substantial differences between the target population and the sample which may have introduced bias into the results. Other countries, such as South Korea, England and Ireland, have also experienced questionable movements in PISA scores over time, potentially due to sampling issues (Eivers, 2010; Micklewright et al., 2012). Other

criticisms of PISA include potential bias introduced by cross-national and cross-cultural differences in the translation, interpretation and understanding of the test questions (El Masri et al., 2016; Kankaraš & Moors, 2014).

However, perhaps the most controversial element of PISA is the scaling model (i.e. how a country's PISA scores are derived from students' responses to the test questions). This consists of two core components: An Item Response Theory (IRT) model and a latent regression model. Together they form the so-called 'conditioning model', from which estimates of students' achievement in reading, mathematics and science are derived (OECD, 2014a). This is a complex, multi-step procedure; one which has been criticised for being opaque (Goldstein, 2017) and is not well understood outside the psychometric community.

This scepticism about the PISA scaling model has been shown to be warranted by some academic research. For instance, Wuttke (2007) has challenged the assumption that each PISA subject can be measured via a single unidimensional latent trait. He also questioned whether all test items really function the same across all populations in such a diverse sample. Fernandez-Cano (2016) questioned PISA's historic use of Rasch over other possible IRT models, and the fact that certain characteristics of test questions (e.g. different response formats, position effects) are not accounted for. Kreiner and Christensen (2014) made a similar criticism, providing evidence of general misfit of test questions within the PISA scaling model and evidence of significant differential item functioning. They consequently concluded that cross-country comparisons of educational achievement in PISA should be handled with great care (Kreiner & Christensen, 2014). Meanwhile, Rutkowski (2014) illustrated how systematic error within background variables could bias subpopulation estimates of students' achievement. In contrast, Jerrim et al. (2018) suggest that relative differences between

OECD countries remain largely unchanged after a series of alterations to the IRT component of the PISA scaling model were made.

However, one element of the PISA scaling model that has been subject to less scrutiny – despite it being the subject of quite some criticism and confusion – is the role that background information about students plays in the derivation of PISA scores. Specifically, students' responses to questionnaire items (e.g. their socio-economic background, their attitudes towards school etc.) are used in conjunction with their responses to the PISA test questions to generate the PISA 'plausible values' (PISA estimates of students' academic achievement). In particular, all students are assigned these plausible values in a given subject (e.g. reading) based upon how well they performed in test questions covering other subjects (e.g. mathematics and science) and the responses they provided to the background questionnaires (e.g. their gender, socio-economic status etc). Importantly, they receive these plausible values in each subject regardless of whether they took any test questions in that subject or not.

For those outside the psychometric community, the idea that such background data plays a role in the generation of PISA scores is difficult to understand. However, it is argued that, as PISA is only interested in achievement at the aggregate (e.g. country) level, and not in the achievement of individual pupils, then this should not bias the results. At the same time, the use of background data in the scaling model (in theory) brings two important advantages. First, if this is not done, then attenuation bias may be introduced when looking at the covariation between PISA scores and background characteristics (Mislevy, 1991; Mislevy et al., 1992). Second, by conditioning upon pupils' background characteristics, the precision of population estimates should be enhanced (e.g. smaller standard errors in average PISA scores; van Rijn, 2018). On the downside, this adds substantial complexity to the generation of PISA scores.

While conditioning upon background characteristics is a key part of the production of PISA scores, only two out of nineteen chapters of the PISA 2012 technical report are dedicated to the computation of plausible values (OECD, 2014a). This highlights the lack of examination of the topic, which is also evidence from the scarcity of research conducted on this matter (most of the literature cited above focuses upon the IRT part of the scaling model). For instance, do cross-country comparisons of PISA scores change depending upon if (and how) the conditioning model is specified? Does it really bring the supposed benefits that motivates its use?

This paper aims to answer such questions about the conditioning model used in PISA and fill the gap in the literature. It begins by investigating how closely the PISA plausible values can be reproduced using publicly available documentation about the procedures used. We then compute alternative plausible values using different variants of the conditioning model. Results from using the full conditioning model are then compared to those using only basic parts of the model, to those using no conditioning model at all. This, in turn, allows us to establish whether (a) cross-country comparisons of PISA scores change depending upon the conditioning model used and (b) whether the theoretical benefits of conditioning upon background data are empirically observed in this setting.

The results from this analysis lead us to four key conclusions. First, while the publicly available information provided by the OECD allow close replication of the plausible values in the major domain (mathematics in the PISA 2012 data we use), replications for the minor domains (especially reading) are less successful. The OECD, consequently, need to be much more transparent about exactly how PISA plausible values for the minor domains have been derived – and particularly about the precise specification of the conditioning model. Second, while the specification of the

conditioning model has little influence upon the PISA ranking within the major domain (mathematics), there is an impact in some of the minor domains (particularly reading). Third, there is evidence that the specification of the conditioning model can have substantial, but not necessarily predictable, impacts upon important measures of educational inequality. Finally, we find no evidence that population estimates (e.g. average PISA scores) become more precise (i.e. standard errors are smaller) when a complex conditioning model is used. Actually, the opposite holds true (standard errors inflate rather than deflate).

Methods

Data

In this paper, we use PISA 2012 data to illustrate how PISA scores are derived. Generally, PISA aims to compare the mathematics, reading and science skills of 15-year-olds across countries. To achieve this aim, nationally representative samples of 15-year-olds who are enrolled in at least grade 7 in an educational institution are drawn (OECD, 2014a, p. 66). PISA 2012 encompasses 478,413 students in 64 countries and economies (Cyprus excluded).

Test design

As time is a limiting factor in educational assessment, PISA uses a rotated test design. This means that, in PISA 2012, students were randomly assigned to complete one of 13 different test booklets. Each of these booklets contained four out of 13 possible 'item clusters' (groups of questions). As mathematics was the focus of PISA 2012, seven of the 13 item clusters were about this subject, with three of the clusters about science and

three clusters about reading[1]. All booklets contained at least one mathematics item cluster, but only five of 13 booklets included questions in each of reading, mathematics, and science. In other words, only around 40% of students answered questions in all three core PISA domains (OECD, 2014a, pp. 30, 31). Therefore, complex techniques (IRT and latent regression) are used to impute data in domains where students have not answered any test questions (e.g. reading) from how they performed upon test questions in other domains (e.g. mathematics and science) and their background characteristics (e.g. gender, socio-economic status, attitudes towards mathematics). See OECD (2014a, pp. 145, 146) for further details.

A unique feature of PISA 2012 (which did not occur in prior or subsequent PISA rounds) was that rotation was also used for the student background questionnaire, resulting in three different versions. These questionnaires shared a common core component, while also including a rotated part that differed. Hence, while some information (e.g. gender, language and parental education) is available for all students, some other background data are only available for a subset (OECD, 2014a, p. 58). In addition to the mandatory questionnaires and domains (student and school questionnaires and the mathematics, reading and science test), countries could also administer some optional elements of PISA. This included parental, educational career and information communication technology questionnaires as well as additional assessments in digital reading, computer-based mathematics, financial literacy and problem solving (OECD, 2014a, pp. 22, 259, 260; see Appendix A for more details).

A summary of how PISA plausible values are generated

Using students' responses to the test questions and questionnaire, the survey organisers follow five main steps to compute the PISA scale scores (plausible values) (see chapter

9 and 12, especially pp. 159, 253, 254 of OECD, 2014a).

- First, for each core domain (reading, mathematics, and science) the item difficulties are determined using a common sample[2] via IRT. These are then fixed for all later stages.
- Second, responses to the background questionnaires are recoded for each country. These are then used as 'conditioning variables' in subsequent steps.
- Third, student achievement distributions are estimated. This is done separately in each country via a combination of IRT and latent regression (known in the psychometric literature as a conditioning model). In short, both students' responses to the test questions and the responses provided to the background questionnaires are used to estimate student's achievement in each subject. A simplified illustration of the model used can be found in Figure 1. However, rather than providing a single point estimate of the achievement for each student, a conditional achievement distribution is generated.
- Fourth, for each student, five plausible values are randomly drawn from this distribution. Within the literature, these are viewed as 'imputations' for unobserved (latent) student achievement (Mislevy, 1991).
- Finally, these plausible values are transformed by common item equating to the PISA scale. This final element facilitates comparisons of PISA scores over time.

The focus of this paper is the role of the conditioning model detailed in the third bullet point above[3].

<Figure 1>

Why are background variables used within the construction of PISA scores?

Despite conditioning models having now been used for decades in large-scale international assessments, the PISA technical reports provide little rationale for their use; it has simply been described as a 'natural extension' of IRT (OECD, 2014a, p. 145). In a nutshell, they are essentially an application of Rubin's (1987) well-known multiple imputation (MI) methodology applied to IRT, treating students' latent abilities as an extreme form of missing data. The motivation for their use hence closely follows the rationale put forward in the MI literature; it is necessary to include background variables in the estimation of students' latent abilities in order to (a) facilitate unbiased estimations of group differences (e.g. difference in achievement between boys and girls)[4] – see (Mislevy, 1991; Mislevy et al., 1992) and (b) reduce uncertainty in measurement (van Rijn, 2018).

This rationale shows why it is important that PISA (and other international surveys) use a conditioning model. However, as noted by Wu (2005), it is important that this model is correctly specified. Otherwise, bias might be introduced. This not only holds true for average PISA scores (the subject of much attention), but also measures of educational inequality and differences between key sub-groups (e.g. how gender and migrant-native student gaps compare across countries). Indeed, while there are strong theoretical arguments for PISA's use of a conditioning model, the substantial complexity it introduces has meant it has thus far not been closely scrutinised (Goldstein, 2017).

Replication of the PISA methodology

In order to investigate how the specification of the conditioning model influences PISA results, we begin by attempting to replicate the PISA methodology of creating plausible

values as closely as possible. Following the formulas and annotation used within the OECD technical reports (OECD, 2014a, pp. 144–146), let:

- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)$ denote the latent variable of the D domains,
- $f_\theta(\boldsymbol{\theta}; \boldsymbol{\alpha})$ be the density of the of the latent variable $\boldsymbol{\theta}$,
- $\boldsymbol{\alpha} = (\mu, \sigma^2)$ denote the parameters of the density for a unidimensional latent variable and $\boldsymbol{\alpha} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a multidimensional,
- $\boldsymbol{Y_n}$ denote a vector of u values (e.g. background characteristics) for student n and
- $\boldsymbol{\beta}$ be a vector of regression coefficients.

The following paragraphs focus on the core part of the conditioning model as defined in PISA; we adopt the IRT model and its response vector as it described within the technical report. Assuming that the density of a certain latent achievement (θ) follows a normal distribution with $N(\mu, \sigma^2)$, as done within PISA, then the density function becomes[5]:

$$f_\theta(\theta_i; \boldsymbol{\alpha}) = (2\pi\sigma^2)^{-\frac{1}{2}} exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2}\right].$$

In the above, no conditioning model has been applied. Now, let's assume that students from different sub-populations have different abilities. The density function above now needs to be tweaked to reflect this (which is done via the conditioning model). While the variance of the density is fixed, the mean μ is replaced with the regression model estimate $\boldsymbol{Y_n'\beta}$. As a result, the latent variable is now represented through $\theta_n = \boldsymbol{Y_n'\beta} + \varepsilon_n$**,** with the independent error term having zero mean and being normally distributed. Note that $\boldsymbol{Y_n}$ can consist of several different background

characteristics which researchers may want to relate to student achievement within secondary analyses.

If we plug this regression into the density function, we end up with the following conditioning model:

$$f_\theta(\theta_n; \boldsymbol{Y_n}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(\theta_n - \boldsymbol{Y_n'\beta})'(\theta_n - \boldsymbol{Y_n'\beta})\right].$$

This can be generalised to facilitate multidimensional latent variable estimation (e.g. the estimation in PISA of students' reading, science, and mathematics abilities) using a multivariate normal distribution with respective parameters:

$$f_\theta(\boldsymbol{\theta_n}; \boldsymbol{w_n}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta_n} - \boldsymbol{\gamma w_n})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta_n} - \boldsymbol{\gamma w_n})\right].$$

In this case $\boldsymbol{\gamma}$ is a matrix of the regression coefficients with the different dimensions, $\boldsymbol{\Sigma}$ is the variance-covariance matrix for the D dimensions and $\boldsymbol{w_n}$ is the vector of fixed variables equivalent to $\boldsymbol{Y_n}$ in the unidimensional case.

Empirically, we apply this approach to the PISA 2012 data as described in Appendix B.

How are student background data incorporated into the plausible values?

As stated above, the conditioning variables are a vital part of the conditioning model. In PISA 2012, all variables from the background questionnaires are recoded, pre-processed[6] and then used as conditioning variables ($\boldsymbol{Y_n}$). Within the conditioning model, each background variable is treated as either (OECD, 2014a, p. 157):

- A direct regressor. These are added straight to $\boldsymbol{Y_n}$ without any further processing, just recoding. Only the following variables are direct regressors:

gender, school ID, grade, mothers and fathers socio-economic index and booklet IDs[7]. These variables are available for all students in the PISA conditioning model.

- An indirect regressor. The remaining (vast majority) of background variables are recoded or pre-processed in one or two out of three ways: (a) combined into preliminary questionnaire indices; (b) dummy-coded if categorical or (c) centred and a dummy variable added for missing information if numerical[8]. A principal component analysis is then conducted on these recoded variables, with as many components retained as necessary to explain 95% of the variance. The retained components are then included in the vector of conditioning variables $\boldsymbol{Y_n}$. According to the official documentation, no imputation, or other approaches to dealing with the large amounts of missing background data (due to the rotated questionnaire design) were applied. The conditioning variables $\boldsymbol{Y_n}$ are computed separately by country and may therefore vary (e.g. available information and number of principal components that were retained per country).

Variations of the conditioning model

After trying to reproduce the plausible values, we then alter how the conditioning variables are used in the PISA scaling process to examine how the specification of the conditioning model affects cross-country comparisons of PISA scores.

To achieve this goal, the conditioning variables are divided into three groups: (a) school-level direct regressors (contrast codes for school ID), (b) individual-level direct regressors (all remaining contrast codes) and (c) indirect regressors. Using different combinations of the above, we generate eight alternative sets of plausible values, each based upon a different specification of the conditioning model:

(0) No conditioning variables (i.e. no conditioning model at all)

(1) School direct regressors only

(2) Individual direct regressors only

(3) Indirect regressors only

(4) All direct regressors (school + individual)

(5) School direct regressors and indirect regressors

(6) Individual direct regressors and indirect regressors

(7) All regressors (as used in PISA).

This enables us to analyse how the specification of the conditioning model affects cross-country comparisons of PISA scores.

All computations and analyses within this paper are done within R (R Core Team, 2019) using the 'TAM' (Robitzsch et al., 2018) and 'intsvy' (Caro & Biecek, 2017) packages. Further details about the computational procedures (both the replication and altering the conditioning variables) can be found in Appendix C. For the comparisons and analyses of the produced plausible values, we accounted for the sample design by using Balance-Repeated-Replication (BRR) weights in combination with the final student weight.

Analyses

Average scores

Initial replication

Figure 2 illustrates the relationship (at the country level) between our self-computed country average PISA scores and the 'official' OECD scores. The upper panel refers to our plausible value computation without conditioning (i.e. background variables have

not been included in the conditioning model). The lower panel is where the full conditioning model (including all variables stated in the PISA 2012 technical report) has been used.

<Figure 2>

Our replication of the PISA plausible values has succeeded to different degrees. The correlation between our country averages and the 'official' country averages is very good for the major domain (mathematics) where correlations are above 0.998. Similar results hold for science (one of the minor domains). Although there is slightly more variation between the official country average science scores and our replicated values, the cross-country correlation in the results is still strong; the Pearson correlation is .996 with full conditioning and .998 without. In other words, in these two domains, the impact of conditioning upon the results is trivial.

The results for reading (the other minor domain) are, however, more of a concern. In the upper panel, when no conditioning is applied, our country averages closely replicate the official OECD scores (Pearson correlation = .997). This changes in the bottom panel once we condition upon background data, with the correlation falling slightly to .965, leading to many countries experiencing an important change to their results. For instance, at the extreme, the average reading score in Chile increases from 441 to 479 (i.e. by more than 0.3 standard deviations), while it falls in the Netherlands from 511 to 490 (i.e. a drop of around 0.2 of an international standard deviation). Indeed, when conditioning upon background characteristics, our estimates of average reading scores in lower performing countries have a slight tendency to be higher than the official results, while our average reading scores for high performing countries tend to be slightly lower.

Given these results, from this point forward, we focus mainly upon findings for reading in the main text. Some general comparisons between the domains can be found in Appendix D. Full results for all three domains separately can be found in Appendix E (mathematics), F (science) and G (reading).

The impact of conditioning

To illustrate the possible impact of conditioning on average reading scores, we focus on the comparison of our self-computed plausible values with and without conditioning. This can be found in Figure 3. The lines depict the effect that conditioning has on country average reading scores.

<Figure 3>

In general, average reading scores within most countries decline when conditioning is applied (triangular markers in Figure 3 tend to be lower than the circular markers), with only 13 out of 62 countries experiencing an increase. Indeed, as Figure 3 demonstrates, the impact of conditioning in low-performing countries is relatively small (the circle and triangular markers tend to sit on top of each other) while in middle-to-high performing countries the impact of conditioning seems larger (the circle and triangular markers are farer apart). Yet, there are some expectations in lower-performing countries like Chile and Colombia, which also experience a substantial impact on their average scores. In terms of the often-cited PISA 'country-rankings', conditioning has relatively little impact upon the composition of the top and bottom performing groups with some exceptions. It does, however, lead to important changes around the middle, where country averages are close to each other and changes due to model specification occur in different magnitudes and directions. For instance, Israel drops 13 places (from 25th to 38th) while Portugal rises 15 places (form 29th to 14th).

What part of the conditioning model leads to this difference? The next part of the analysis compares results using different specifications of the conditioning model, focusing upon three different subsets of conditioning variables: (a) school direct regressors (i.e. contrast codes for each school); (b) individual direct regressors (e.g. gender, socio-economic status) and (c) indirect regressors (i.e. the rest of the background questionnaire variables that have been reduced into a set of principal components).

Table 1 shows the average country reading scores of the OECD countries for different specifications of the conditioning model. The shading should be read vertically (within conditioning model specification) with green (red) cells indicating higher (lower) average scores.

<Table 1>

While most countries stay roughly in the same area of relative achievement, there remains variation and changes in ranking between the different model specifications. For some countries, the relative position changes quite substantially depending upon specification (e.g. Portugal, Norway, and Chile). For instance, the cross-country correlation between the results with no conditioning (M0) and with any form of conditioning tends to be around 0.78 to 0.85 with exception of M5 (school direct and indirect regressors) with a correlation of 0.92. Likewise, there is variation in the extent of correlation of the different specifications with the full model. No conditioning (M0) and indirect regressors only (M3) show the lowest correlation with 0.85, while the model with the other two components (M4 – school direct and individual direct regressors) reaches correlation of 0.96 with the full conditioning model (M7).This suggests that it is not only the decision of whether to use conditioning that is important, but also the precise specification of the conditioning model.

The average reading scores (and ranking) for selected countries are particularly sensitive to conditioning model specification. For example, the performance for Israel drops substantially when all direct regressors are used as in M4 and M7 (orange cell, corresponding to 30th place). But it displays visibly lighter orange/yellow colour for the other models (between 16th and 19th place for other model with exception of 23rd place for individual direct and indirect regressors) This suggests the selection of conditioning variables can have a significant (and yet unpredictable) impact upon countries' average PISA scores in at least one of the minor domains.

Inequality in PISA scores

While country average scores receive a lot of attention, the data are also used in many other ways. One of the most prominent examples is in cross-country comparisons of educational inequality. For instance, since 2009 PISA dedicates the whole second volume of their international reports towards equity and outcomes, while UNESCO uses PISA for their report on educational inequality (Gromada et al., 2018). Educational inequality using PISA has also been the subject of much academic research (e.g. Oppedisano and Turati (2015) and Gamboa and Waltenberg (2012)). We therefore illustrate in Table 2 how sensitive a widely used measure of educational inequality (the difference between the 90th and 10th percentile) is to different specifications of the conditioning model. Green (red) shading in this table illustrates lower (higher) levels of inequality.

The first key point of note from Table 2 is that conditioning leads to an increase in estimated educational inequality (on average) across OECD countries. Specifically, the average percentile gap rises by 23 points, from 211 with no conditioning to 234

when full conditioning is applied. The gap between the 90th and 10th percentile increases substantially as soon as any conditioning is used.

<Table 2>

Second, the relative position of countries in international comparisons of educational inequality appears more sensitive to the specification of the conditioning model than the average scores. The cross-country correlation between M1-M6 and M7 (full conditioning) generally falls between 0.79 and 0.91. At the same time, none of the specifications shows a particularly high correlation (r between 0.63 and 0.83) with M0 (no conditioning applied). In general, high variation between the different specifications can be seen through the varying colour patterns.

Finally, no clear country patterns can be identified, either in relation to changes in average scores nor concerning changes between model specifications. Norway, for example, has high fluctuation in the level of inequality measure depending on the chosen specification (between 2nd and 24th place). Whereas other countries, such as Turkey, see a rather constant shift as soon as conditioning is applied (4th place without conditioning and between 10th and 15th place as soon as conditioning is applied).

The association between PISA scores and background characteristics

PISA is also often used to compare the performance of groups (e.g. gender, socio-economic status). Yet it is well known that IRT when used in conjunction with rotated test designs can lead to attenuation of such group differences (Mislevy, 1991). One of the main motivations for using conditioning models is to counteract such attenuation bias. We begin by illustrating this issue with respect to gender differences, as this is one of the major group comparisons focused upon within the OECD PISA reports (e.g. 3 of the 14 statements in the 2012 executive summary address gender gaps; OECD, 2014b).

Gender is an individual direct regressor meaning that, once direct regressors have been included in the conditioning model, the potential problem of attenuation bias should be resolved.

For this purpose, we take a closer look at models M0, M2 and M7 to examine how the specification of the conditioning model impacts the gender gap (computed by regressing reading performance upon an indicator whether the student is female). Figure 4 hence illustrates the gender gap in reading using model M0 (no conditioning - circle), M2 (just direct individual regressors including gender - diamond) and M7 (the full model - triangle).

For most countries, the diamond (M2) and triangle (M7) are pointing in the same direction, and for about half of those, they sit on top of each other. This suggests that, in most countries, the gender gap is not sensitive to the exact specification of the conditioning model (once gender has been included as a direct regressor) with a potential small increase or decrease in the full model. There are, nevertheless, some important changes to the results for some individual countries (that are somewhat difficult to explain). Visible differences between M0, M2 and M7 occur in multiple countries. For instance, in Norway and the United Arab Emirates (framed by the two blue boxes) the estimated gender gap from M2, the model including gender, is even more similar to M0, with a large jump in the magnitude of the gender gap in M7. Such changes are perplexing and again suggests that the precise specification of the conditioning model applied can have an impact upon a key aspect of a country's results.

<Figure 4>

Thus far, we have focused upon gender as a 'direct regressor' (meaning it is entered directly into the PISA conditioning model). Yet most background data collected in PISA are 'indirect regressors' - meaning they are only incorporated into the

conditioning model having first been pre-processed using a Principal Component Analysis (recall subsection 'How are student background data incorporated into the plausible values?' in 'Methods' for further details). Investigating whether the relationship between indirect regressors and PISA scores changes depending upon the specification of the conditioning model is hence also of interest. The analyses of migrant status yielded similar results to gender and can be found in Appendix H.

The impact of conditioning upon standard errors

Another goal of conditioning, apart from counteracting attenuation, is higher precision in group estimates (van Rijn, 2018). To conclude this section, we therefore investigate how conditioning affects the standard error of country average scores. Figure 5 provides a boxplot illustrating how the standard error of the mean changes for different specifications of the conditioning model. One would anticipate that the boxplots should move southwards as one moves from left (M0) to right (M7) – as more information is being used about students to derive the plausible values. But this is not the case; standard errors are typically *higher* once conditioning is used. In fact, in mathematics and reading no country had a smaller standard error when full conditioning was used (compared to no conditioning). In science, only four countries (Singapore, Macao, Estonia, and Canada) experienced an increase in precision when full conditioning was applied. However, in general, no substantial benefit can be found for precision from conditioning, with standard errors actually inflating, if anything.

<Figure 5>

Discussion

PISA is an international large-scale assessment which examines the educational

achievement of 15-year-old students across the world. It aims to provide comparable achievement scores in mathematics, reading and science between countries and groups, as well as over time. This has resulted in PISA becoming one of the key studies used for evidence-based education policymaking across the globe. As a tool which can potentially influence many people's lives, it is essential that the statistical foundations that underpins this study are sound. Yet, time and again, criticisms have been made about the opaqueness of PISA's methodology (Goldstein, 2017). Despite this, relatively little research has closely scrutinized key aspects of the PISA scaling model. This includes 'conditioning', where background variables are used in the derivation of the PISA plausible values.

This paper has tried to fill this gap in the literature. Specifically, we have re-estimated PISA 2012 scores for each participating country having altered key aspects of the conditioning model. This includes investigating how key results change when different sets of background variables are used in the PISA conditioning model, and what happens when no conditioning variables are used in the construction of PISA scores at all. We not only document the impact that this has upon average country scores, but also cross-national comparisons of educational inequality (i.e. the spread of achievement) and gaps in performance between different groups (e.g. gender differences).

Our results illustrate how the precise specification of the conditioning model does indeed matter, though the impact this has depends upon both the subject and the statistic of interest. In terms of average scores, results for the major domain can be considered 'robust' (i.e. unaffected by whether/how conditioning variables are used). Yet results for the minor domains are more mixed. Although the specification of the conditioning model has little impact upon cross-country comparisons of average scores

in science, the same is not true for reading where average scores (and, consequently, rankings) change. Rather different results were obtained for educational inequality, where cross-country comparisons in all three domains were sensitive to the specification of the conditioning model. The conditioning model specification was also found to have some impact upon the magnitude of group differences, with particularly big changes observed for gender differences in reading and mathematics in a few countries.

While we believe this study illustrates some important points about the PISA scaling methodology, findings should be interpreted considering its limitations. First, while great effort has been made to replicate the official PISA methodology, there remained some differences between our self-computed plausible values and those provided in the OECD PISA database. Although we believe that the approach, we have taken provides a sufficient basis for the present study, it is not a perfect replicate for what the OECD (and their contractors) have done. Unfortunately, the OECD do not release their code for how they have constructed the PISA scores. To be as open as possible about our own approach (and to allow other researchers to independently scrutinise our findings) we have made freely available the code we have used to produce our results (available from https://github.com/lrzieger/ upon publication).

Second, we focus on the methodology used for one specific PISA cycle (2012). We note that the scaling model (including the conditioning) changed in PISA 2015 and with the introduction of computer adaptive testing in 2018. This means that this paper is not directly applicable to subsequent PISA cycles, though still yields some important lessons learnt. A key direction for future research is hence to replicate our findings for other PISA cycles, both prior and subsequent to 2012, to develop a better understanding of whether the role of the conditioning model on key country-level outcomes has changed over time.

Third, relatedly, we have only developed a limited understand of what drives the changes in the results across the different conditioning model specifications, particularly in reading. In a companion paper (Zieger and Jerrim *forthcoming*) we have undertaken a simulation study to probe this issue further. In this we find it does not seem to matter how the variables included in the conditioning model are pre-processed - what matters more is the variables that are chosen. Moreover, the bias that we have noted in this paper for reading – which we also find in our simulation study – seems to be largely driven by the inclusion of booklet dummy variables within the conditioning model. Finally, any bias which is introduced through the conditioning model specification can be effectively counteracted if all students answer some questions in all domains.

Finally, we did not recompute the scale identification but used the transformation provided within the PISA technical reports. As it is a linear transformation, this could potentially affect the comparability of absolute numbers between the official and our self-computed scores. Yet this issue does not affect relative achievement positions (such as rankings) or the cross-country correlation of results, which are the focus of this paper.

Despite these limitations, the findings of this paper does speak to the psychometric literature on the pros and cons of conditioning. First, as noted by Mislevy (1991), if one does not condition upon background variables in certain test designs – such as those used in PISA 2012 – then this may lead to biased estimates of group differences. Our results demonstrate this empirically, highlighting how gender differences – most notably in reading scores – suffer from severe bias unless conditioning is applied. On the other hand, another theoretical benefit of conditioning is that the extra information contained within the background data will result in more precise estimates (van Rijn 2018). Empirically, however, we find no evidence that this

is the case. Hence the sole motivation for using a conditioning model should be to guard against bias when estimating group differences, rather than in the hope that this will lead to appreciably smaller standard errors.

Given that our results point towards some changes in certain key statistics – such as measures of inequality – depending upon the scaling model used, one may question whether this puts at stake the validity of secondary analysis of PISA (including those conducted by the OECD themselves). While we believe that such statistics should be interpreted with some care, they do nevertheless provide some important information about how educational achievement gaps compare across countries. Our advice is therefore that multiple different sets of plausible value are generated - each using different versions of the conditioning model specification – so that the robustness of any such comparisons can be easily and thoroughly interrogated. This would have the key advantage that such important statistics would continue to be produced, but also that their limitations are more widely appreciated and understood.

We also hope this paper has made a valuable contribution to ongoing debates about PISA's methodology. It adds three key points. First, the technical report is not detailed enough to allow independent researchers to exactly replicate and closely scrutinize the scaling model and its resulting plausible values. The OECD must become more transparent in its methodology and to make its technicalities more digestible – particularly to non-specialized audiences. As part of this, the OECD should commit to publishing the code it uses to produce the plausible values, helping to facilitate greater independent verification of the PISA scores. Second, educationalists and policymakers the world over should note from our findings that, while results from the major domains appear to be quite trustworthy and robust, those for the minor domains should be treated with care. Third, relatedly, we are aware that the OECD is currently consulting in PISA

moving from a three-year to a four-year cycle, with each domain then receiving equal assessment time in each wave. The results from this paper, along with our companion simulation study (Zeiger and Jerrim *forthcoming*) lead us to strongly support this proposal (as the "conditioning" upon background questionnaire items becomes less important when there is more test score information available for a given subject area). Our view is that the findings that we report here are unlikely to have been completely resolved in subsequent PISA cycles, and that the only way that they will be is for all pupils to receive test questions in each of the cognitive domains. Finally, we question PISA's reliability as a valid way to measure educational inequality across countries, given the major impact the conditioning model specification can have upon the results. All the above leads us to plead with the OECD that additional sensitivity analyses around the PISA results must be conducted and be transparently reported with the release of every future cycle.

Acknowledgement. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement no. 765400.

Notes. [1] Two of the mathematics item clusters exist in an 'easy' and a 'standard' version (clusters 6 and 7). Countries with a low expected performance can administer the easy versions instead of the standard versions. This leads to 13 booklets per country in either the easy or standard version, with an overlap of six booklets.

[2] The common sample existed of 500 students from each country, expect for Liechtenstein, which were randomly selected (OECD, 2014a, p. 233).

[3] As a result, the first and final part of the procedure described above will not be directly replicated in this paper. Rather, the officially published numbers (e.g. values of item difficulties) will be used instead.

[4] In the MI literature, it is widely suggested that (in the presence of missing data) the relationship between a variable and the outcome of interest will be attenuated unless that variable is included in the imputation model. This idea is also applied within the conditioning modelling literature, with it being claimed that the relationship between students' background characteristics and their achievement will be attenuated unless that variable is included in the conditioning model.

[5] For the estimation of an IRT model, some assumptions need to be made. There are different approaches to enable the estimation. The approach involving the specification of a density for the latent variables is called the 'marginal approach' and is used in PISA.

[6] By recoding, we mean altering and transforming the format of the variable without changing the meaning or value of the variables (e.g. contrast/dummy-coding of variables). By pre-processing, we mean altering and transforming the values of the variables (e.g. computing a new questionnaire index by averaging multiple variables or using principle components).

[7] The contrast coding for booklets was further tweaked so that the information for students who only answered questions in two domains is based on information from all booklets that have items in a domain (OECD, 2014a, p. 157). Furthermore, the regression coefficients for booklets which covered two of three domains were set to zero for the third domain in the latent regression.

[8] The exact details for all recoding can be found in Annex B in the technical report (OECD, 2014a, pp. 421–431).

References:

Caro, D. H., & Biecek, P. (2017). Intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software*, *81*(1), 1–44. https://doi.org/10.18637/jss.v081.i07

Egelund, N. (2008). The value of international comparative studies of achievement–a Danish perspective. *Assessment in Education: Principles, Policy and Practice*, *15*(3), 245–251. https://doi.org/10.1080/09695940802417400

Eivers, E. (2010). PISA: Issues in implementation and interpretation. *The Irish Journal of Education / Iris Eireannach an Oideachais*, *38*, 94–118.

El Masri, Y. H., Baird, J.-A., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619–634. https://doi.org/10.1080/03054980600976320

Fernandez-Cano, A. (2016). A methodological critique of the PISA evaluations. *Relieve*, *22*(1), 1–16.

Freitas, P., Nunes, L. C., Balcão Reis, A., Seabra, C., & Ferro, A. (2016). Correcting for sample problems in PISA and the improvement in Portuguese students' performance. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 456–472. https://doi.org/10.1080/0969594X.2015.1105784

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review*, *31*(5), 694–708. https://doi.org/10.1016/j.econedurev.2012.05.002

Gillis, S., Polesel, J., & Wu, M. (2016). PISA Data: Raising concerns with its use in policy settings. *The Australian Educational Researcher*, *43*(1), 131–146. https://doi.org/10.1007/s13384-015-0183-2

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

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy*, *24*(1), 23–37. https://doi.org/10.1080/02680930802412669

Gromada, A., Rees, G., Chzhen, Y., & Cuesta, J. (2018). Measuring inequality in children's education in rich countries. *Innocenti Working Papers*. https://doi.org/10.18356/5f90f95e-en

Hopmann, S., Brinek, G., & Retzl, M. (2007). *PISA according to PISA: Does PISA keep what it promises?* (Vol. 6). LIT Verlag.

Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How robust are cross-country comparisons of PISA scores to the scaling model used? *Educational Measurement: Issues and Practice*, *37*(4), 28–39. https://doi.org/10.1111/emip.12211

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381–399. https://doi.org/10.1177/0022022113511297

Kreiner, S., & 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. https://doi.org/10.1007/s11336-013-9347-z

Meyer, H.-D. (2014). The OECD as pivot of the emerging global educational accountability regime: How accountable are the accountants? *Teachers College Record*, *116*(9), 1–20.

Micklewright, J., Schnepf, S. V., & Skinner, C. (2012). Non-response biases in surveys of schoolchildren: The case of the English Programme for International Student Assessment (PISA) samples. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(4), 915–938.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

OECD. (2014a). *PISA 2012 technical report*. OECD.

OECD. (2014b). *What students know and can do: Student performance in mathematics, reading and science* (Rev. ed., Febr. 2014). OECD.

Oppedisano, V., & Turati, G. (2015). What are the causes of educational inequality and of its evolution over time in Europe? *Education Economics*, *23*(1), 3–24. https://doi.org/10.1080/09645292.2012.736475

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation. www.R-project.org

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules* (R package version 3.1-45). https://CRAN.R-project.org/package=TAM

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education*, *27*(2), 115–132. https://doi.org/10.1080/08957347.2014.880440

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Sellar, S., & Lingard, B. (2013). Looking East: Shanghai, PISA 2009 and the reconstitution of reference societies in the global education policy field. *Comparative Education*, *49*(4), 464–485. https://doi.org/10.1080/03050068.2013.770943

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*(4), 387–407. https://doi.org/10.1080/03050060802481413

van Rijn, P. (2018, November 7). *Basic principles of population modelling*. IERI Academy hosted by CARPE, Dublin.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Measurement, Evaluation, and Statistical Analysis*, *31*(2), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Wuttke, J. (2007). Uncertainty and bias in PISA. In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA: Does PISA keep what it promises* (pp. 241–263). LIT Verlag.