CENTRE FOR EDUCATION POLICY
& EQUALISING OPPORTUNITIES

UCL

Conditioning: How background variables can influence PISA scores

Centre for Education Policy and Equalising Opportunities (CEPEO)

Laura Zieger, John Jerrim, Jake Anders & Nikki Shure

Working Paper No. 20-09

April 2020

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- PISA data is widely used as a basis for education policymaking. Yet, few people realise that students background characteristics are used in the creation of PISA scores, and why this is done.
- This is at least partly due to the fact that the methodology used in PISA is complex and opaquely communicated. In this paper we first replicated how PISA scores are created to demonstrate this process. We then systematically alter how the background variables (such as student characteristics) are used in the computation of PISA scores to investigate how this affects the results.
- While countries' mean achievement is robust for the major domain, different specifications in how PISA scores are generated were found to lead to important changes for one of the minor domains (reading).
- This sensitivity of PISA results to the precise methodology used is even more pronounced when we look at measures of inequality. Changes to how background variables are used lead to large changes in terms of how countries compare for educational inequality.
- All in all, we show precise choice of the statistical model underlying the creation of PISA scores can make a fundamental difference to the results. We therefore urge the OECD to conduct and publish more sensitivity analyses around how the results are produced. Cross-country comparisons of educational inequality based upon PISA should be treated with particular caution.

Why does this matter?

PISA has been widely used to compare inequality in educational achievement across countries. This paper suggests these comparisons should be regarded critically.

Conditioning: How background variables can influence PISA scores

Abstract: The Programme for International Student Assessment (PISA) is an international large-scale assessment which examines the educational achievement of 15-year-old students across the world. It has long become one of the key studies for evidence-based education policymaking across the globe. As result, PISA results and the methodology that they are based on should be robust, open and transparent. Yet, PISA receives significant criticism for its scaling model and the opaqueness in communicating it. One particular point of concern is the so-called "conditioning model", where background variables are used in the derivation of student achievement scores. The aim of this paper is to investigate this part of the scaling model and the impact it has upon the final scores. This includes varying the background variables of the conditioning model systematically and analysing the impact that this has on multiple measures. Our key finding is that the exact specification of the conditioning model matters and has substantial impact on average scores in some of the minor PISA domains (namely reading). It also has a major impact upon cross-national comparisons of educational inequality.

Acknowledgement: Laura Zieger, John Jerrim, Jake Anders and Nikki Shure are part of the European Training Network OCCAM. 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.

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, with reference societies (such as Finland) having emerged (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 the most controversy. As discussed by Rutkowski and Rutkowski (2016) and others (Gillis, Polesel, & Wu, 2016; S. Hopmann, Brinek, & Retzl, 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, Schnepf, & Skinner, 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, Baird, & Graesser, 2016; Kankaraš & Moors, 2014).

However, perhaps the most controversial element of PISA (and the area that has received most criticism) is the scaling model used (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. A seminal paper by 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 (i.e. a lack of measurement invariance across countries). 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 (provided within the background questionnaires) 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" (the closest thing in PISA to estimates of students' academic achievement). 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, Beaton, Kaplan, & Sheehan, 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, leading to the criticisms that it is opaque.

While conditioning upon background characteristics is a key (if poorly understood) part of the production of PISA scores, relatively little existing research has been 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 (smaller standard errors and more accurate estimates of covariation with background characteristics)? Or does it simply add a great deal of complexity (and fuel criticisms of PISA lacking transparency) for little discernible gain?

This paper aims to answer such questions about the so-called "conditioning model" used in PISA. 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 (achievement estimates) 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 scores (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 a big impact in some of the minor domains (particularly reading). In other words, different versions of the conditioning model can lead to rather different country-level PISA scores. Third, 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). Finally, there is evidence that the specification of the conditioning model can have substantial, but not necessarily predictable, impacts upon important measures of educational inequality.

This then leads us to two key recommendations. First, as others have previously suggested, the scaling procedure used in PISA is not sufficiently transparent to facilitate exact replication of the results by independent researchers. The technical reports supplied by the OECD do not contain sufficient detail about the procedures used (let alone in a language suitable outside of a highly specialised field) and should therefore be extended. Second, the specification of the conditioning model can lead to non-trivial changes to average PISA scores, particularly within minor domains. Based upon this evidence, we conclude that the OECD should publish more sensitivity analyses around the conditioning model and make more detailed information about their methodology publicly available.

Methods

Data

In this paper, we use PISA 2012 data to illustrate score computation in PISA. Generally, PISA aims to compare the mathematics, reading and science skills of 15-year-olds between 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). A two-stage stratified sample design is used. In the first stage, at least 150 schools are sampled per country with probability proportional to school size. Subsequently, 35 students per school are randomly sampled. In some countries, larger samples are drawn in order to facilitate sub-population (within-country) comparisons (OECD, 2014b, p. 256). The average school and student response rates after replacement are 98% and 92%, though there are substantial differences between countries.

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] Consequently, 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). The survey organisers therefore use complex techniques (item-response theory and latent regression) 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, enjoyment of school). 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. Specifically, there were three different versions of the student questionnaire, to which students were also randomly assigned. 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

[1] Each cluster contained 30 minutes of test material. Two of the mathematic item clusters exist in an easy and a standard version (mathematics item cluster 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.

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 (EC) and information communication technology (ICT) 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 details). The additional domains were computer-based assessment, while the core domains were paper-based.

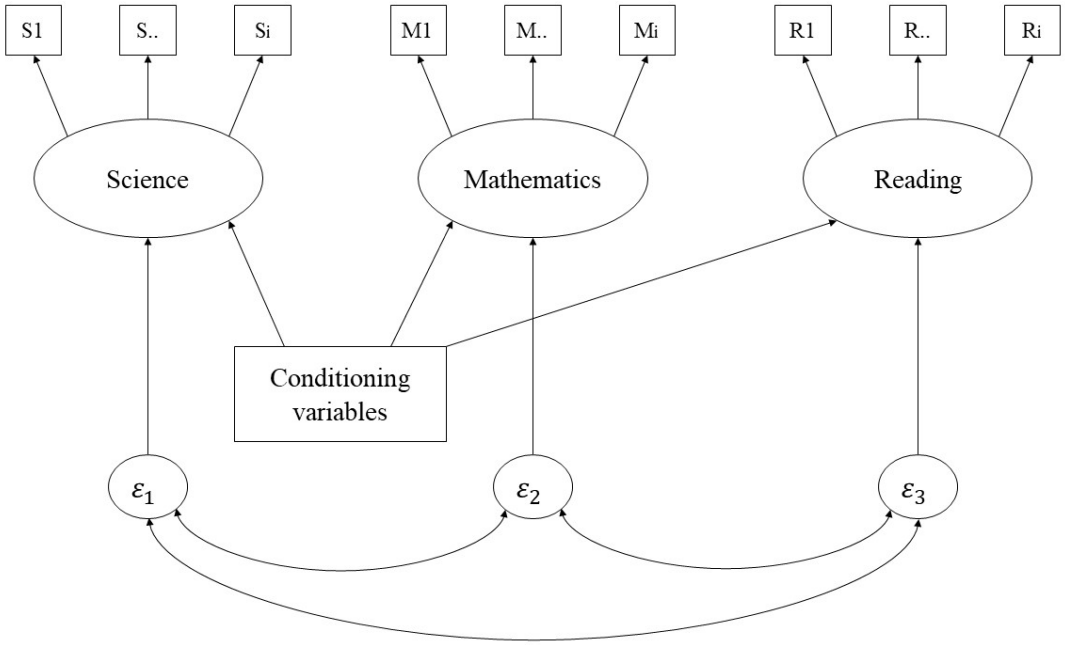
A summary of how PISA scale scores (plausible values) are generated

Using students' responses to the test questions and questionnaire items to which they were randomly assigned, the survey organisers follow five main steps to compute the PISA scale scores (plausible values) found within the publicly available PISA database (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 item response theory (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. Further details about this part of the procedure will be discussed below.
- Third, student achievement distributions are estimated. This is done separately in each country via a combination of item response theory (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. This distribution reflects, for each student, the uncertainty we have in their estimated reading, science and mathematics ability.
- 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.

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

The focus of this paper is the role of the "conditioning model" (i.e. the use of school and student background data) detailed in the third bullet point above[3].



Note. Squares refer to observed variables, ovals to latent variables and circles to error terms. S.., M.., and R.. refer to students' responses to PISA test questions, where i is the number of items in the domain. Curved lines connecting errors indicate correlated errors.

Figure 1. A simplified illustration of the PISA scaling model used to generate the plausible values

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).

[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 (i.e. there will be downward bias in the estimated coefficient) 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.

The idea behind the first of these points is best explained with a simplified example. Imagine a rotated assessment design where only half of the students receive reading questions, but all receive mathematics questions. Now assume that female students achieve 10 achievement points more in reading than their male counterparts, but that there is no gender difference in mathematics. If a standard IRT model is applied (without conditioning upon gender), students who did not answer the reading questions would be assigned a reading score based solely upon their responses to the mathematics questions. Consequently, for the part of the sample that were given only mathematics questions, girls would be assigned the same reading scores as boys. This would in turn mean that, were we to estimate gender differences in reading achievement across the whole sample, we would find a difference of just five test points rather than 10 (i.e. there would be attenuation bias affecting the results). When using complex rotated test designs, estimates of such group differences hence need to be adjusted in order to produce unbiased results. Within PISA, this is likely to be particularly important for the minor domains, where there are large amounts of "missing data".

This simple example illustrates why it is important that PISA (and other international surveys) use a conditioning model. However, as noted by Rutkowski (2014) and Wu (2005), it is important that this model is correctly specified. Otherwise, bias might be introduced. At a minimum, it is vital that thorough investigations are undertaken to consider how PISA results might change if a different conditioning model is used. 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). The aim of this paper is to fill this gap in the literature.

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:

- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$ 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_i) follows a normal distribution with $N(\mu, \sigma^2)$, as done within PISA, then the density function becomes[5]:

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

In the above, no conditioning model has been applied. Now, let's assume that students from different sub-populations (e.g. boys and girls) 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 stays the same, the mean μ is replaced with the regression model estimate $\boldsymbol{Y_n'\beta}$. As a result, the latent variable is now represented through $\theta_{in} = \boldsymbol{Y_n'\beta} + \varepsilon_n$, with the independent error term having zero mean and being normally distributed. Note that $\boldsymbol{Y_n}$ can consist of several different background characteristics (e.g. gender, grade, parental education, attitudes towards school, young people's self-efficacy) 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:

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

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.

[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.

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 deviation contrast coding. Only the following handful of variables are direct regressors: gender, school ID, grade, mothers and fathers socio-economic index and booklet IDs[7]. These variables are therefore available for all students in the PISA conditioning model[8].
- An indirect regressor. The remaining (vast majority) of background variables are recoded in one 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[9]. A principal component analysis (PCA) 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. in terms of the number of components that were retained). For each country, all available information was used[10].

Analytical aim of this paper

While the PISA technical reports contain a lot of information, only two of the nineteen chapters are dedicated to the computation of the plausible values. It therefore lacks the finer details about the computational procedures. We nevertheless try to reproduce the published plausible values as closely

[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 categorical variables: instead of having one variable existing of all different categories, we have an indicator for the categories (-1 due to not adding a reference category indicator) which is 1 if the student answered in that category, -1 if in the reference category was selected or zero if neither). 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). Further details on the recoding and pre-processing used in PISA 2012 can be found in the technical report (OECD, 2014a, pp. 157, 421–431).

[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.

[8] This is true even with the questionnaire rotation used in PISA 2012, as questions capturing this information was seen by all students.

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

[10] For example, Germany administered the parental questionnaire. This meant that more items were included in the PCA for the computation of indirect regressors in Germany than in most other countries.

as possible. 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 will generate eight alternative sets of plausible values, each based upon a different specification of the conditioning model. These eight alternatives can be summarised as follows:

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).

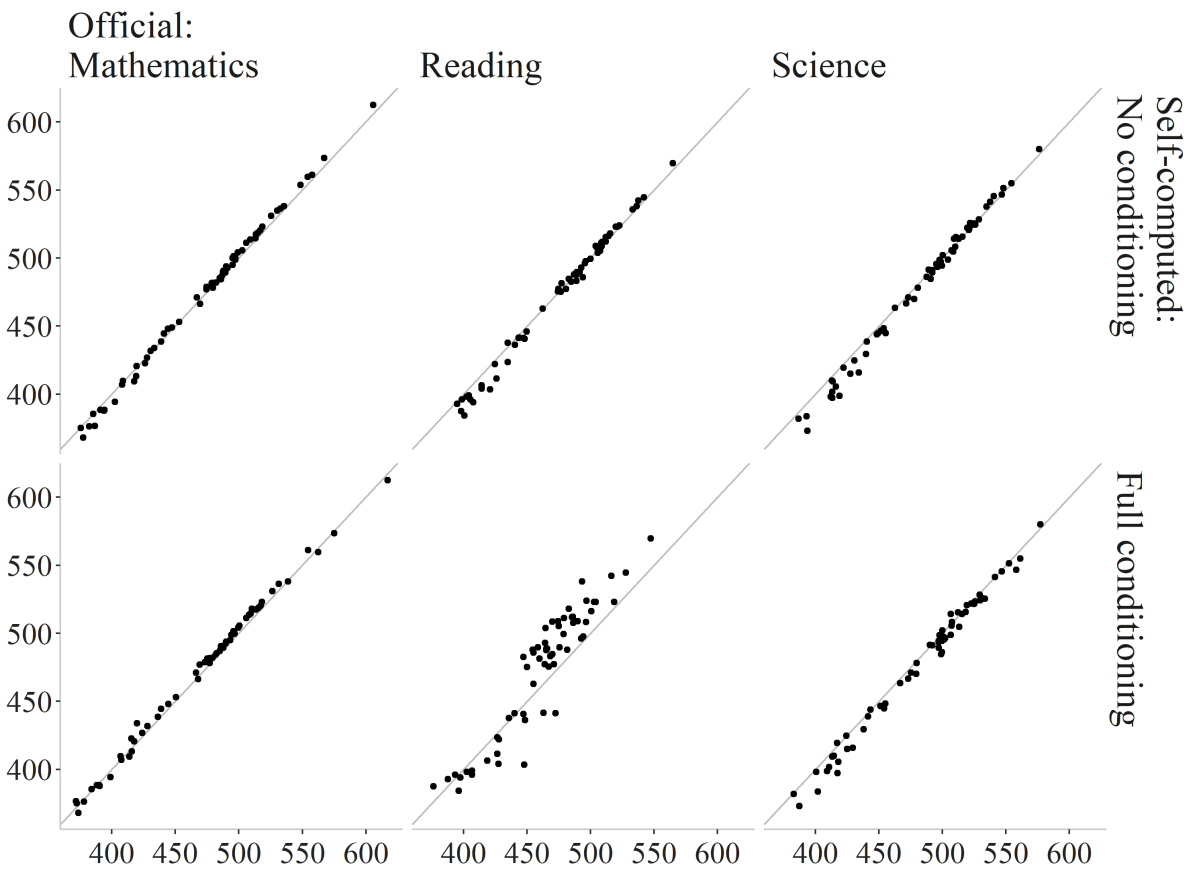
This enables us to analyse how the specification of the conditioning model affects cross-country comparisons of PISA scores.

All computations and analyses within this paper are done within R (R Core Team, 2019) using the 'TAM' (Robitzsch, Kiefer, & Wu, 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 BRR weights in combination with the final student weight.

Results

Average scores

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.



Notes: The 'official' country average scores are plotted along the horizonal axis and our self-computed scores along the vertical axis. The upper panel refer to results where no conditioning upon background characteristics has been applied. The lower panel is where the full conditional model (as described in the PISA 2012 report) has been applied. The 45-degree line is where these two values are equal. The Person correlations, starting in the top-left hand graph and working right, are .999, .997, .997, .998, .936 and .995.

Figure 2. Countries' average PISA scores. Official versus self-computed scores

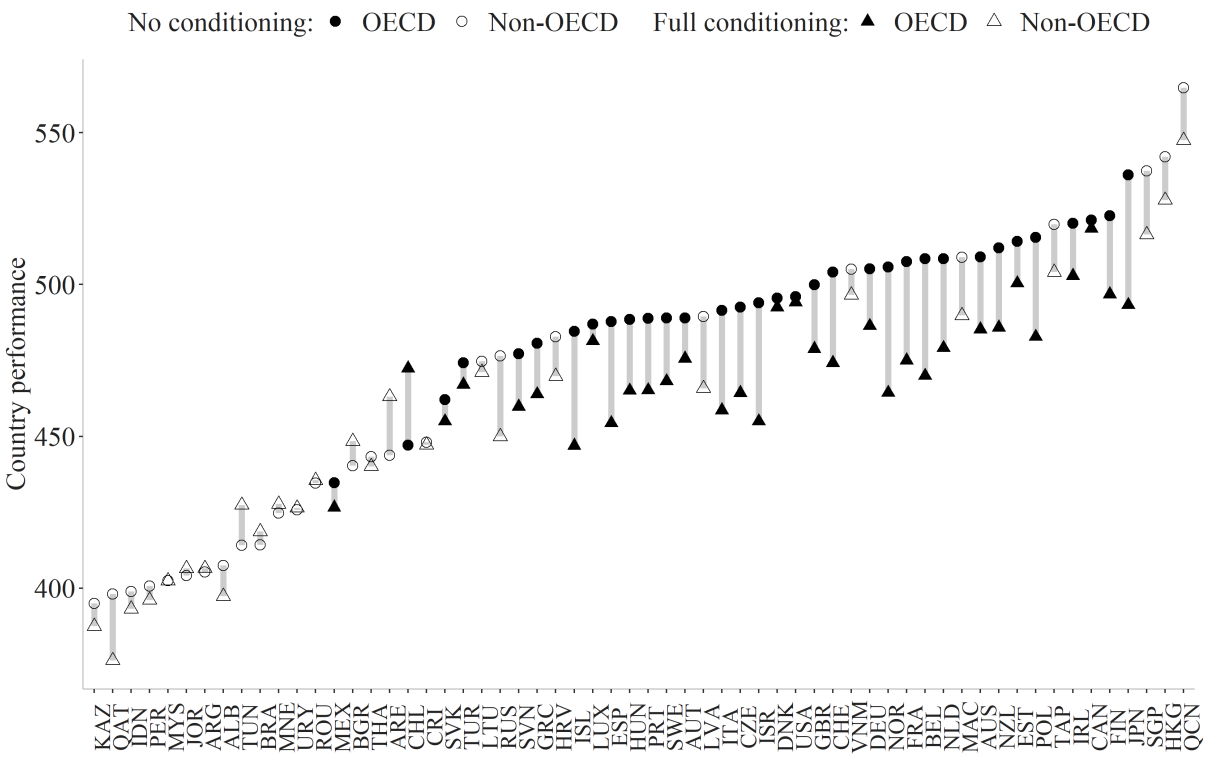
Our replication of the PISA plausible values has succeeded to different degrees. The correlation between our country averages and the 'official' country averages is very good for the major domain (mathematics) where the correlation is above 0.998 (regardless of whether conditioning is used). This illustrates two key points: first, consistent with Jerrim et al. (2018), independent replication of country average major domain scores is feasible and, second, cross-country comparisons of country averages in the major domain do not vary depending upon whether conditioning modelling is used.

Similar results hold for science (one of the minor domains). Although there is slightly more variation between the official country average and our replicated values, the cross-country correlation in the results is still strong; the Pearson correlation is .995 with conditioning (lower panel) and .997 without (upper panel). This demonstrates that country average scores in one of the minor domains (science) can also be replicated and are not affected by whether conditioning is used.

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. Specifically, the correlation between our replicated country averages and the official PISA average reading scores falls to .936, with many individual countries experiencing an important change to their results. For instance, at the extreme, the average reading score in Chile increases from 441 to 472 (i.e. by around 0.3 standard deviations or roughly a year of additional schooling), while in Japan it falls from 538 to 493 (i.e. a drop of almost half an international standard deviation). Indeed, when conditioning upon background characteristics, our estimates of average reading scores in lower performing countries tend to be higher than the official results, while our average reading scores for high performing countries tend to be lower. This highlights an important finding; whether one uses background data in the construction of PISA scores can lead to large changes in the comparative performance of countries within at least one of the minor domains. In other words, the method used to derive the PISA scores can have a significant impact upon country rankings, independent upon how students responded to the test questions.

Given these results, from this point forward, we focus upon findings for reading in the main text. Full results for all three domains can be found in Appendix D (mathematics), E (science) and F (reading). We also note in our discussion where findings differ across the three domains.

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.



Notes: Triangles provide estimates without conditioning and circles with conditioning. Solid markers are OECD countries and hollow markers non-OECD countries.

Figure 3. Country average reading scores with and without conditioning

In general, average reading scores within most countries decline when conditioning is applied, with only 10 out of 60 countries experiencing an increase. Interestingly, OECD countries (solid markers) experience an average drop of 21 points in reading scores when we apply conditioning, which is much larger than the (on average) six-point decline in non-OECD countries (hollow markers). 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 when looking at the left-hand side of Figure 2) while in middle-to-high performing countries the impact of conditioning seems much larger (the circle and triangular markers are quite far apart when looking at the right-hand side of Figure 3). Moreover, only one of the 10 countries with a positive increase in reading scores after conditioning are members of the OECD: Chile (+25 points). In terms of the often-cited PISA 'country-rankings', conditioning has relatively little impact upon the composition of the top and bottom performing groups. It does, however, lead to important changes around the middle. For instance, Norway drops 15 places (from 17th to 33rd) while the Chile rises 19 places (form 43rd to 24th).

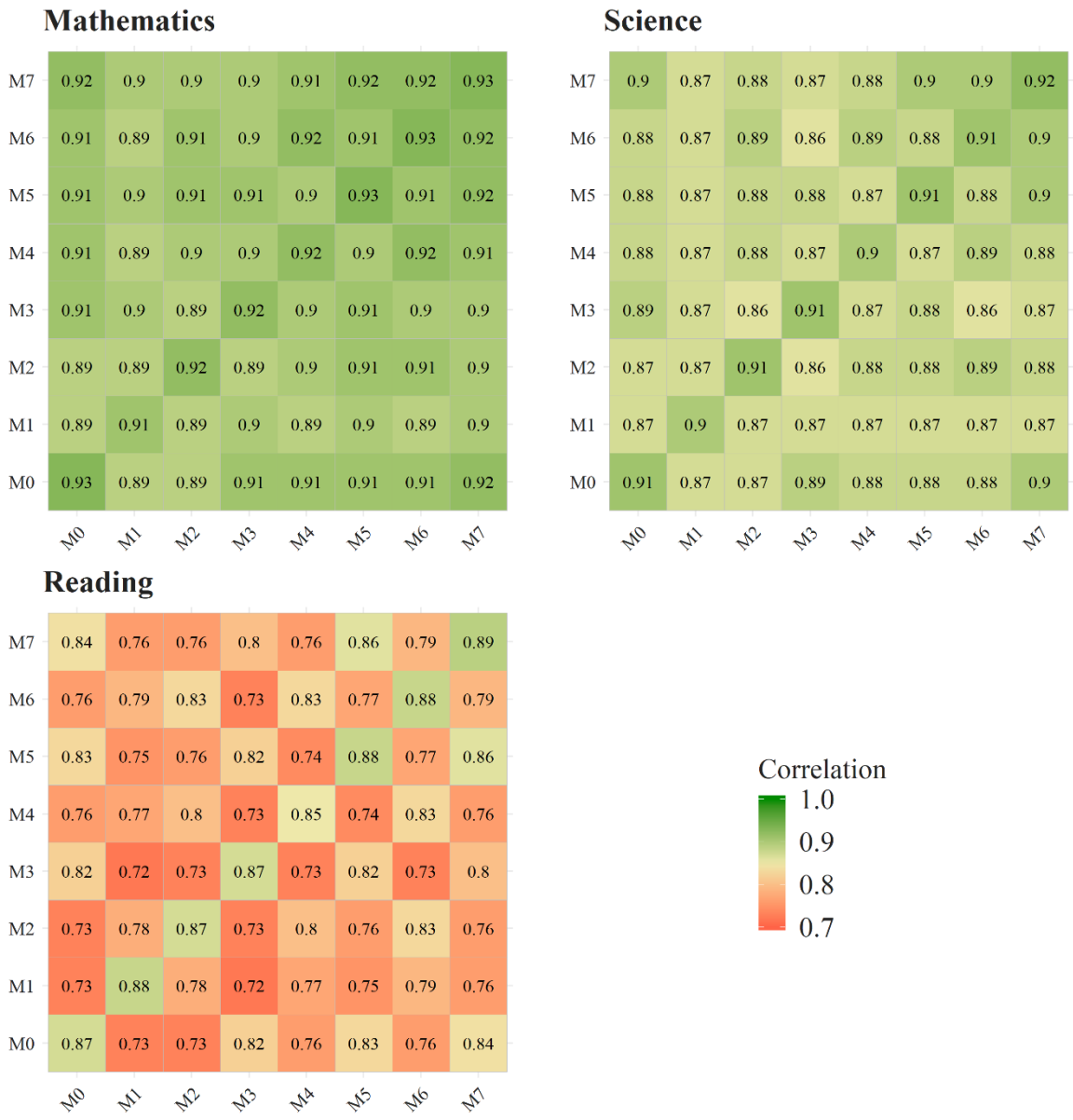
What part of the conditioning model leads to this difference? The next part of the analysis compares results using different specifications of the conditioning model, focusing upon three different subsets of conditioning variables: (a) school direct regressors (i.e. contrast codes for each school); (b) 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).

Figure 4 displays the correlation between the plausible values (at the individual level) using different specifications of the conditioning model. The greener a square is, the closer the correlation is to one. On the other hand, red shading denotes a correlation of 0.7 (around the minimum we observe across any model).

Two points come to attention. First, the shading clearly illustrates that the correlation varies between the domains. As expected, the results for mathematics (the major domain) have the strongest correlations across different conditioning model specifications. While the correlations for science are slightly lower, those for reading are particularly low (as illustrated by the predominance of red squares). This highlights how, although the precise specification of the conditioning model has little impact upon the results in the major domain of mathematics, it has important implications in the minor domains (particularly reading). As the minor domains have a lot fewer test questions in the PISA test design than the major domain, and given the correlation between mathematics and reading achievement is likely to be substantially lower than the correlation between mathematics and science achievement, this finding makes sense.

Second, these findings are reinforced when looking at the diagonals in Figure 4. The correlations sit between 0.91 and 0.93 in mathematics, 0.90 and 0.92 in science and 0.85 and 0.89 in reading. As plausible values incorporate uncertainty about individual achievement, higher correlations between the plausible values created using different conditioning models partially reflect the greater certainty in measurement. Reading hence has lower correlations than mathematics and science due to the extra uncertainty in the results for this domain.



Notes. The correlations are based on individual-level plausible values across all countries. The colour scale ranges from $r = .7$ (red) to $r = 1$ (green). M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Figure 4. Correlations of the individual-level plausible values in mathematics, reading and science with different specifications of the conditioning model

Table 1 goes one step further and 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. The rows at the bottom provide the OECD average/median and the correlation of results across different model specifications.

Table 1. Variation in estimated average PISA reading scores by conditioning model specification. OECD countries.

Country	M0	M1	M2	M3	M4	M5	M6	M7
Japan	536	526	501	523	493	527	493	493
South Korea	533	542	498	-	500	-	-	-
Finland	523	524	487	524	491	524	492	497
Canada	521	544	521	530	519	537	521	518
Ireland	520	518	518	524	506	523	506	503
Poland	515	528	479	515	481	517	480	483
Estonia	514	531	514	536	511	534	516	500
New Zealand	512	512	480	512	481	512	483	486
Australia	509	501	478	497	484	505	474	485
Belgium	509	502	463	506	467	509	460	470
Netherlands	509	509	474	509	476	509	475	479
France	508	511	467	500	473	510	471	475
Norway	506	491	458	462	461	493	462	464
Germany	505	509	492	509	488	514	490	486
Switzerland	504	508	466	506	477	507	475	474
United Kingdom	500	499	472	499	474	498	476	479
Denmark	496	489	473	506	484	502	479	492
USA	496	510	502	503	500	508	496	494
Israel	494	507	443	498	447	504	448	455
Czech Republic	493	489	459	490	462	488	460	464
Italy	491	491	449	490	453	490	453	459
Austria	489	491	455	488	460	493	469	476
Hungary	489	483	456	489	458	487	458	465
Portugal	489	526	452	501	452	501	463	465
Sweden	489	520	447	506	450	503	491	468
Spain	488	489	447	488	449	488	452	454
Luxemburg	487	487	453	487	453	486	476	481
Iceland	485	485	445	484	446	484	446	447
Greece	481	481	448	481	464	481	463	464
Slovenia	477	485	449	477	464	492	466	460
Turkey	474	474	458	474	481	474	469	467
Slovak Republic	462	478	452	471	440	478	444	455
Chile	447	489	481	494	463	477	480	472
Mexico	435	432	414	432	422	432	430	427
OECD average	**497**	**502**	**469**	**497**	**471**	**500**	**473**	**474**
OECD median	**496**	**502**	**465**	**499**	**470**	**502**	**474**	**474**
Correlation with M0	**1.00**	**0.83**	**0.70**	**0.80**	**0.74**	**0.91**	**0.63**	**0.74**
Correlation with M7	**0.74**	**0.76**	**0.93**	**0.84**	**0.93**	**0.84**	**0.93**	**1.00**

Notes: Figures illustrate how average PISA reading scores vary depending upon the specification of the conditioning models. Results for non-OECD countries reported in Table F.1. Green shading indicates higher scores relative to other countries and red cells lower scores. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. South Korea is missing scores due to computational difficulties.

Relatively few countries (e.g. New Zealand and Czech Republic) maintain a stable position in the PISA reading rankings across all different specifications of the conditioning model; most countries relative reading scores change depending upon the model specification. For instance, the cross-country correlation between the results with no conditioning (M0) and with any form of conditioning tends to be around 0.70 to 0.91. Likewise, there are sometimes substantial differences between using a relatively sparse conditioning model (e.g. M1 or M3) and the full conditioning model (M7). Furthermore, it is striking that all specifications including the individual direct regressors (M2, M4 and M6) are the only ones with correlation above 0.85 (0.93 in all three cases) with the full conditioning model (M7, which also includes individual direct regressors). The individual direct regressors thus appear to be a dominant factor in the conditioning model. 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. It is also noteworthy how the OECD average reading score changes non-trivially between model specifications, from a minimum of 469 (individual direct regressors) to a maximum of 502 (school direct regressors).

The average reading scores (and ranking) for selected countries are particularly sensitive to conditioning model specification. Take, for example, Portugal. This country has a relatively high performance (a green shaded cell, corresponding to 6th place) when only school direct regressors are used. But it then lights up in orange for all other specifications (15th and 18th place for indirect regressors (and direct school regressors) and otherwise between 22nd and 28th place). Other countries with very large changes in performance depending upon conditioning model specification include Chile, Norway, Sweden, Israel, Belgium and the United States. 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 PISA scores receive a lot of attention, the data is also used in many other ways. One of the most prominent is in cross-country comparisons of educational inequality; e.g. since 2009 PISA dedicates the whole second volume of their international reports towards equity and outcomes, and UNESCO uses PISA data for their report on educational inequality (Gromada, Rees, Chzhen, & Cuesta, 2018) as well as in research such as 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 28 points, from 212 with no conditioning to 240 when full conditioning is applied. The gap between the 90th and 10th percentile increases substantially as soon as any conditioning is used.

Table 2. Estimates of inequality in PISA reading scores across countries by specification of the conditioning model (P90 – P10 gaps). OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	163	193	200	194	197	195	199	197
Chile	178	137	199	130	201	153	172	207
Estonia	181	173	194	143	203	168	221	212
Ireland	192	171	238	183	199	185	207	198
Turkey	192	217	237	218	220	221	224	225
Denmark	193	190	183	134	205	194	212	211
Poland	198	241	204	209	218	211	216	221
Spain	199	225	241	227	240	227	237	237
Czech Republic	202	218	238	221	228	219	238	227
Canada	205	217	246	190	257	200	254	253
Switzerland	206	236	237	233	240	234	237	242
USA	209	213	207	175	191	171	185	187
Austria	210	209	221	235	206	230	236	240
Germany	210	200	223	204	244	221	230	244
Hungary	210	193	173	201	199	199	193	212
Netherlands	214	237	235	240	241	239	242	242
Slovenia	214	229	228	240	265	205	264	251
Finland	215	241	253	236	253	237	253	249
Portugal	215	215	252	193	246	210	239	261
United Kingdom	216	244	242	239	245	243	243	244
Italy	216	247	266	247	266	247	265	263
Iceland	218	245	246	245	245	246	249	249
Greece	221	250	274	250	269	249	268	268
Norway	221	207	151	168	166	233	191	205
Japan	222	269	259	274	245	262	240	245
Australia	224	239	233	240	253	257	233	254
Belgium	230	209	253	222	289	225	261	255
Sweden	230	228	293	231	282	233	284	270
Israel	239	219	257	251	290	230	294	275
New Zealand	239	266	282	267	282	269	280	278
Slovakia	240	262	260	255	252	264	260	258
France	241	247	242	259	260	255	258	272
Luxemburg	241	269	285	270	285	270	276	272
OECD average	**212**	**223**	**235**	**219**	**239**	**224**	**238**	**240**
OECD median	**214**	**225**	**238**	**231**	**245**	**230**	**239**	**244**
Correlation with M0	**1.00**	**0.71**	**0.58**	**0.71**	**0.69**	**0.77**	**0.72**	**0.78**
Correlation with M7	**0.78**	**0.69**	**0.81**	**0.77**	**0.93**	**0.75**	**0.93**	**1.00**

Notes: Figures illustrate how the difference between the 90th and 10th percentile of PISA reading scores changes depending upon the specification of the conditioning model. Results for non-OECD countries reported in Table F.2. Green shading indicates less inequality in reading scores relative to other countries and red cells greater inequality. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

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 0.7 and 0.95, illustrating how some countries experience an important change to their results, depending upon which model is used. At the same time, none of the specifications shows a particularly high correlation (r between 0.58 and 0.77) with M0 (no conditioning applied). Moreover, the models with the individual direct regressors and at least one other component (M4, M6 and M7) exhibit a particularly high correlation (0.93 as opposed to 0.81 which is the next highest and belongs to the individual direct regressors only) and a rather similar colour pattern. This showcases, especially against the correlation of 0.75 between M5 and M7, that the individual direct regressors play an influential part in the conditioning model.

Finally, the countries whose average scores are most sensitive to conditioning model specification are not necessarily the countries that are also the most sensitive in terms of educational inequality. For example, educational inequality in Chile remains quite stable in Table 2, despite experiencing large swings in its average score (recall Table 1). Now, it seems other countries are impacted more. Norway, Sweden, Canada, Portugal, Japan and Belgium are prominent examples, where estimates of educational inequality are highly sensitive to the specification of the conditioning model. Indeed, Canada swings from below the OECD average when no conditioning is applied (205 versus 212) to becoming a country where educational inequality appears to be rather high (e.g. 253 versus 240 in model M7). Differences in country results with and without any conditioning can also be extreme. Take Norway, for example. This country has comparatively low levels of inequality when conditioning is applied (1st to 8th position with the exception of a 19th place for direct school + indirect regressors), but relatively high inequality when no conditioning is used (23rd). The results for Belgium, on the other hand, fluctuate (between 32nd and 9th place) depending on the chosen specification.

When examining the corresponding tables in mathematics and science (Tables D.1 and D.3 for mathematics and Tables E.1 and E.3 for science), it becomes obvious that the specification of the conditioning model also has substantial influence upon estimates of educational inequality in both other domains. In other words, unlike the results for average scores (where the issue was isolated to reading), estimates of educational inequality are affected across all three domains. This stresses a key point – that the OECD should conduct (and report results from) a much wider array of sensitivity analyses around the specification of the conditioning model. This is particularly true for results in the minor domains – and for measures of educational inequality - where the impact seems to be greatest.

The association between PISA scores and background characteristics

PISA is also often used (including by the OECD) to compare the performance of groups (e.g. gender, socio-economic status, language). But 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 one of the individual direct regressors meaning that, once direct regressors have been included in the conditioning model, the potential problem of attenuation bias should be resolved. However, little research has previously considered how the precise specification of the conditioning model affects cross-country comparisons of group (e.g. gender) differences – which this paper adds to the literature.

Figure 5 illustrates the estimated gender gap across all three domains with and without full conditioning applied (this has been computed by regressing reading performance upon an indicator of whether the student is female). The 45-degree line marks where the gender gap is the same whether conditioning is applied or not. For reading and mathematics, the magnitude of gender differences clearly increases once conditioning has been used (i.e. the data points – countries – are further away from the 45-degree line). Although the points for science are closer to the 45-degree line, Figure 5 nevertheless highlights the general point (already well established in the literature) that failing to include a given factor in the conditioning model can lead to attenuation bias in the results (Mislevy, 1991).

The gender gap differs in magnitude and direction depending upon the domain. In reading, girls perform better than boys independent of the specification of the conditioning model, though the gender gap gets noticeably bigger when conditioning is used (the average gender gap increases from 14 to 33 points). In mathematics, before conditioning is applied, there is (on average across countries) no gender gap (0 points). Yet, when conditioning is applied, boys achieve average mathematics scores 6 points higher than girls.[11] The gender gap is more concentrated in science, with no obvious change occurring when conditioning is used. Only in some – but not all – countries is there evidence of attenuation when conditioning is not used in the scaling model.

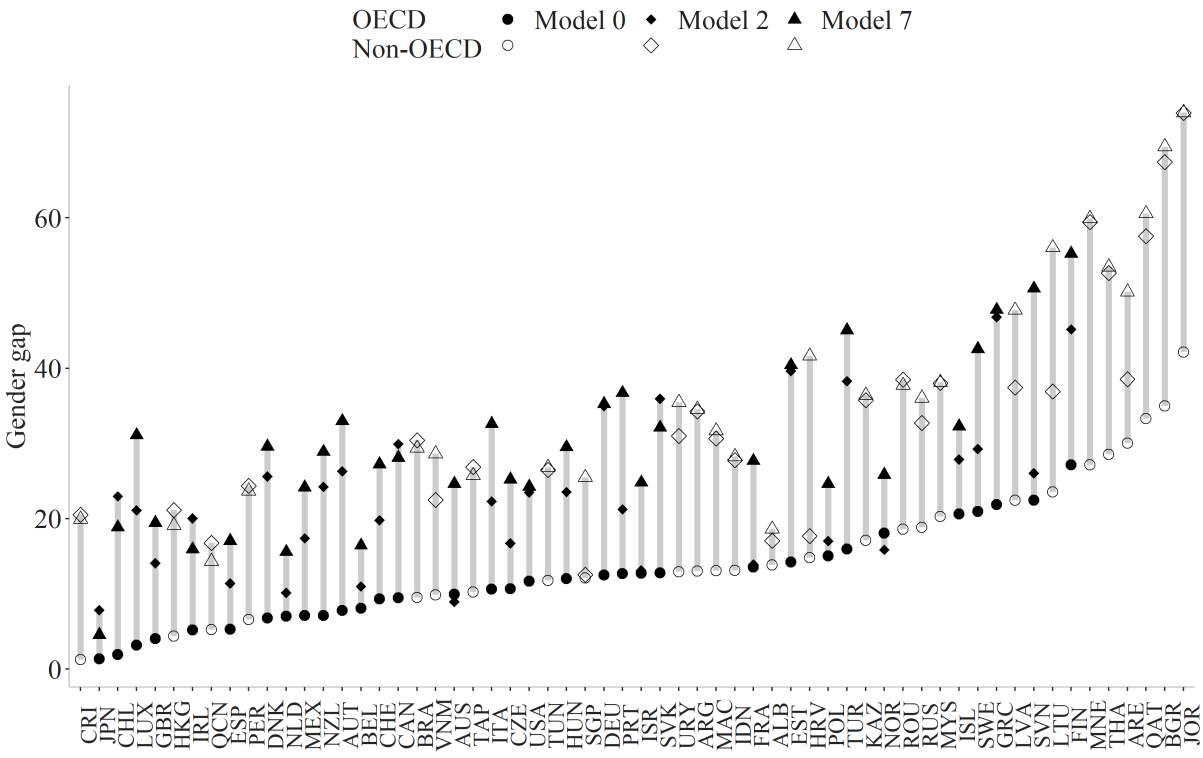
[11] Interestingly, almost all points are below the 45-degree line for mathematics, even the ones with values above zero without conditioning. This means that the mathematics gender gap shifts in favour of boys but is not necessarily moving away further from zero. As a result, attenuation can still be observed in some cases. Finland, for example, has a gender difference of 9 points without conditioning, but only a gender gap of 2 points with full conditioning.

Notes: The gender gaps when using no conditioning are plotted along the horizontal axis and those when using full conditioning along the vertical axis. The 45-degree line is where these two values are equal. The country-level Person correlations, starting left and working right, are $r = .944$, $r = .908$ and $r = .966$.

Figure 5. Country gender gap in mathematics, reading and science with and without conditioning

Next, we take a closer look at models M0, M2 and M7 to further examine how the specification of the conditioning model impacts the gender gap. Figure 6 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 a third 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 by the full model. There are, nevertheless, some important changes to the results for individual countries (that are somewhat difficult to explain). For instance, in Australia, Israel, France, Poland, Slovenia, Norway and Singapore the estimated gender gap from M0 (no conditioning) and M2 (just individual direct regressors) are similar. Yet there is a large jump in the magnitude of the gender gap in M7 (full conditioning model applied). Indeed, in Australia, the gender gap decreases to by nine points when applying M2 (-1 point) but there is a jump in the other direction when using model M7 (+15 points). Such a change in results is 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.



Notes: Circles provide estimates without conditioning, diamonds for conditioning only with individual direct regressors and triangles for full conditioning. Solid markers denote OECD countries and hollow markers non-OECD countries.

Figure 6. Country reading gender gap without conditioning (model 0), just with individual direct regressor incl. gender (model 2) and with full conditioning (model 7)

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 Principle Component Analysis (recall subsection 'How are student background data incorporated into the plausible values?' in 'Data & 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 results from such an analysis are presented in Figure 7, focusing upon migrant status (one of the most widely used contextual variables from PISA that is an indirect regressor in the conditioning model). This shows us how the reading gap between native and migrant students changes between M0 (no conditioning), M3 (just indirect regressors – as captured within the retained principal components) and M7 (the full conditioning model). The key finding from this graph is that the three symbols usually sit on top of each other. In other words, for most countries, it does not matter which conditioning model is used (or whether conditioning is used at all) – you generally get the same result (and indeed the

average gap remains -22 points in M3 and M7 and only changes in the second decimal place). Yet there are again some important exceptions to this finding, most notably Norway with a migrant-native reading gap of -43 points under M0, -1 point under M3 and -19 points under M7. Other countries with large variation in migrant-native achievement gaps tend to have very small proportions of migrant students in the PISA sample, such as Bulgaria (0.4%), Peru (0.5%), Poland (0.2%), Romania (0.1%) and Thailand (0.5%). In Norway, on the other hand, around one-in-ten students are migrants – meaning the fluctuation in the results for this country are unlikely to be due to the small sample size.

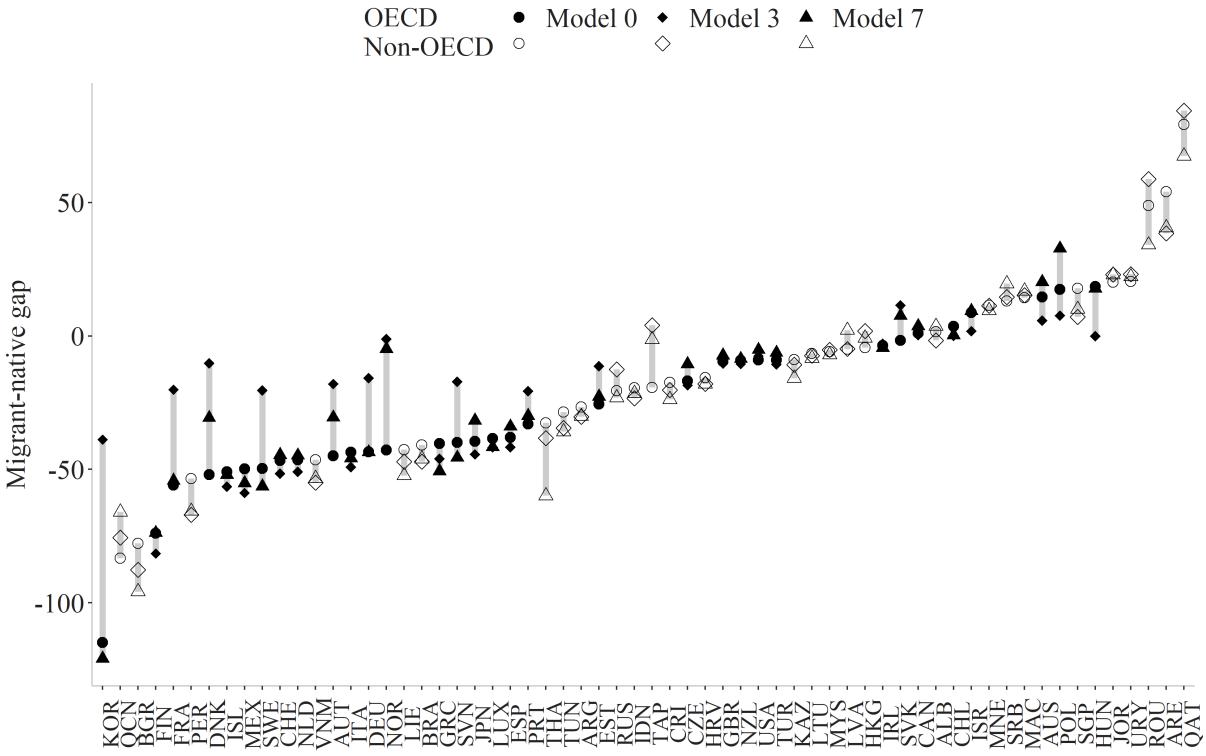


Figure 7. Country reading gap between migrant and native students without conditioning (M0), with indirect regressors (migration status was pre-processed) in conditioning (M3) and with full conditioning (M7)

One might be tempted to conclude from this that it suggests that the PISA results for migrant gaps are generally robust to conditioning model specification. However, an alternative explanation could be that migration status has not been sufficiently represented within the principal components that form the individual indirect regressors. Would the magnitude of the migrant-native gaps change if migrant status was included as direct regressor in the conditioning model instead? We explore this issue in Appendix G, where two further versions of the conditioning model were computed:

- Model M2 was altered to also include migrant status[12] as a direct regressor (M2a)

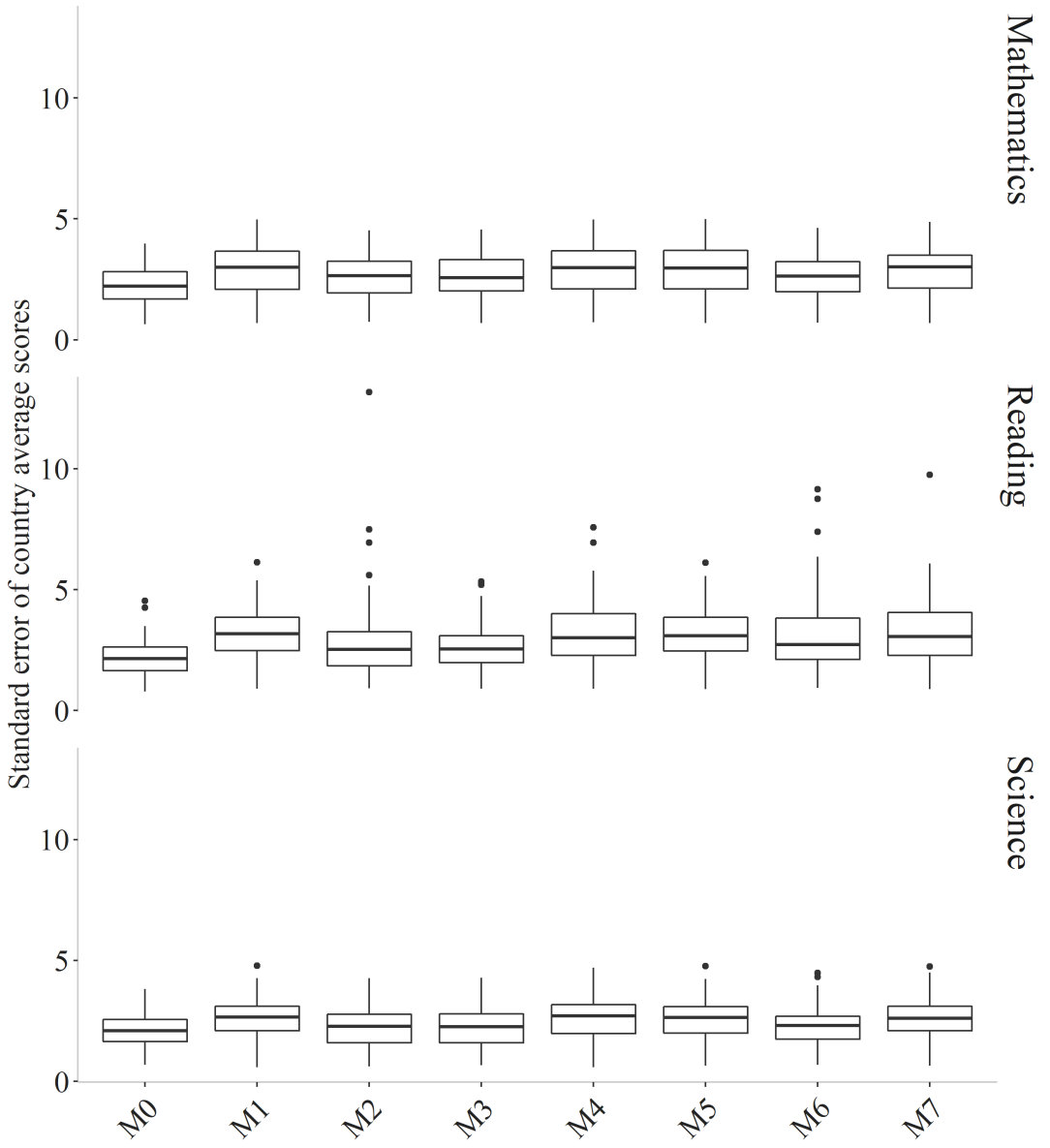
[12] Pre-computed variable 'IMMIG' in the PISA 2012 data set.

- Model M7, the full conditioning model, was re-estimated having included migrant status as a direct regressor, rather than being included within the indirect regressor PCs (M7a)

These models allow us to assess whether including a variable as a direct (rather than indirect) regressor changes the results. In summary, we find that making this change has relatively little impact upon the substantive results. At least in the case of migrant status, including this variable only as an indirect regressor seems to be sufficient.

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 8 provides a boxplot illustrating how the standard error of the mean changes for different specifications of the conditioning model (each country is counted as one data point within each box plot). One would anticipate that the boxplots should move southwards as one moves from left (M0 – no conditioning) to right (M7 – full conditioning). But this is not the case; standard errors are typically *higher* once conditioning is used. In fact, in mathematics no country had a smaller standard error when full conditioning was used (compared to no conditioning). In reading, only one country (Montenegro) experienced an increase in precision when full conditioning was applied, while four countries did in science (Singapore, Macao, Taipei and Estonia). However, in general, no substantial benefit can be found for precision from conditioning, with standard errors actually inflating, if anything.



Notes. The boxplots show the standard errors of the country average score of different countries. M0-M7 denote different specifications of the conditioning model. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning

Figure 8. Boxplots of standard errors of country average scores in mathematics, reading and science with different specifications of the conditioning model

Conclusions

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 level 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. are 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, countries positions in the PISA ranking) change a lot. 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.

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. To be as open as possible about our 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/). We now urge the OECD to do the same.

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 will likely do so again 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. 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, we 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, make the code used to produce PISA scores open-access and to make its technicalities more digestible to non-specialized audiences. 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. 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 (with all code used to produce the results made open-access to allow independent replication of the results). Unless this is done, then sceptics are only right to question whether the PISA results can really be trusted.

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). New York, NY: 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). Vienna, Austria: 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*. Paris: OECD.

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

Pan, J., & Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*, *51*(12), 5765–5775.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342–366. https://doi.org/10.3102/10769986024004342

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

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: 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). *Basic principles of population modelling*. Presented at the 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). Vienna, Austria: LIT Verlag.

Appendices

Appendix A. Which countries participated to what extent in PISA 2012?

As explained in the main body of the paper, countries only had to administer the core domains as well as the student and school questionnaire. Furthermore, countries could opt to administer various additional domains and/or questionnaires. Table A.1 shows the extent of the countries' participation and their sample size. In PISA 2012, 44 countries also administered collaborative problem solving (PS) and 32 countries digital reading and mathematics (DRM). In terms of questionnaires, 23 administered the educational career (EC), 42 the information communication technology (ICT) and only 11 the parental questionnaire. The PISA scaling model uses all available information for a country.

Table A.1. Overview of countries participating in PISA 2012 in the different domains and questionnaires as well as their sample size in the core domains.

			Domain		Questionnaire			
Country	**Abbreviation**	**Sample size**	**PS**	**DRM**	**Parental**	**ICT**	**EC**	**Easier booklets?**
Albania	ALB	4743						
United Arab Emirates	ARE	11500	X	X				X
Argentina	ARG	5908						X
Australia	AUS	14481	X	X		X	X	
Austria	AUT	4755	X	X		X	X	
Belgium	BEL	8597	X	X	X	X	X	
Bulgaria	BGR	5282	X					X
Brazil	BRA	19204	(X)	(X)				X
Canada	CAN	21544	X	X		X		
Switzerland	CHE	11229					X	
Chile	CHL	6856	X	X	X		X	X
Colombia	COL	9073	X	X				X
Costa Rica	CRI	4602					X	X
Czech Republic	CZE	5327	X				X	
Germany	DEU	5001	X	X	X	X	X	
Denmark	DNK	7481	X	X		X	X	
Spain	ESP	25313	(X)	(X)			X	
Estonia	EST	4779	X	X			X	
Finland	FIN	8829	X			X	X	
France	FRA	4613	X	X				
United Kingdom	GBR	12659	(X)					

Greece	GRC	5125					X		
Hong Kong (China)	HKG	4670	X	X	X	X	X		
Croatia	HRV	5008	X		X	X	X		
Hungary	HUN	4810	X	X		X	X		
Indonesia	IDN	5622							
Ireland	IRL	5016	X	X		X	X		
Iceland	ISL	3508					X		
Israel	ISR	5055	X	X			X		
Italy	ITA	31073	(X)	(X)	X	X	X		
Jordan	JOR	7038					X		X
Japan	JPN	6351	X	X			X		
Kazakhstan	KAZ	5808							X
South Korea	KOR	5033	X	X	X	X	X		
Liechtenstein	LIE	293					X		
Lithuania	LTU	4618							
Luxemburg	LUX	5258				X			
Latvia	LVA	4306				X	X		
Macao (China)	MAC	5335	X	X	X	X	X		
Mexico	MEX	33806			X		X		X
Montenegro	MNE	4744	X						
Malaysia	MYS	5197	X						
Netherlands	NLD	4460	X				X		
Norway	NOR	4686	X	X			X		
New Zealand	NZL	4291					X		
Peru	PER	6035							X
Poland	POL	4607	X	X			X		
Portugal	PRT	5722	X	X	X	X	X		
Qatar	QAT	10966							
Shanghai (China)	QCN	5177	X	X		X	X		
Romania	ROU	5074							X
Russian Federation	RUS	5231	X	X			X		
Singapore	SGP	5546	X	X		X	X		
Serbia	SRB	4684	X			X	X		X
Slovak Republic	SVK	4678	X	X		X	X		
Slovenia	SVN	5911	X	X		X	X		
Sweden	SWE	4736	X	X			X		

Chinese Taipei	TAP	6046	X	X		X	
Thailand	THA	6606					
Tunisia	TUN	4407					X
Turkey	TUR	4848	X			X	
Uruguay	URY	5315	X			X	X
United States of America	USA	4978	X	X			
Viet Nam	VNM	4959					X

Notes: Countries in parentheses participated in the additional domains only with a fraction of their sample size, e.g. only one state in the country. In this paper, we do not consider them as administrating this domain.

Appendix B. How does the PISA scaling model take into account different domains and questionnaires being used in different countries?

Not all countries administer all the PISA test domains and questionnaires (e.g. in only a small number of countries is the parental questionnaire collected). As a result, the precise specification of the PISA conditioning model differs between countries (depending upon the extent of their participation). We try to illustrate the subtle differences using Figures B.1 and B.2. This illustrates the following steps:

- Step 1. Item difficulty computation. This step is always the same, regardless of the number of PISA questionnaires and cognitive domains that a country has chosen to conduct. This step is always conducted separately for each domain and is always based upon a common data set encompassing all countries.

However, after this initial step, computations are then conducted separately by country.

- Step 2. Preparation of conditioning variables. This is based on all available background questionnaires for a country and independent of any domain. See the description provided in the section entitled 'How are student background data incorporated into the plausible values?' for further details.
- Step 3. Estimation of student scores. What happens in the third step depends upon the domains of PISA a country participates in (with the exception of financial literacy). If only the core domains are tested, a joint IRT and latent regression model is used for the three domains, where the item difficulties are fixed at the value from step 1[13] (see Figure B.1). Figure B.2 stresses how this step is split into two sub-steps if either (a) problem solving and/or (b) digital reading and mathematics were administered as well. In countries that tested students in these additional subjects, the regression coefficients of the conditioning variables for the core domains are fixed, based upon a joint model consisting of only paper reading, science and mathematics items. This is because "CBA [computer-based assessment] reporting scale cannot influence the PISA paper-based assessment" (OECD, 2014a, p. 157). For those countries, the first joint model is only used to retrieve the regression coefficients for the core domains, but a second joint model is used for the final student achievement estimation. In this second model, all available domains are used (e.g. problem solving can influence science), but additionally the regression coefficients for the core domains are fixed at the values from first joint model.
- Step 4. Plausible values are drawn from the individual conditional achievement distribution, which is based on the final model within each country. It involves all available cognitive domains, whether this is just the three core domains (reading, mathematics and science), four

[13] The published item difficulties are used in our case.

domains (the three score domains plus problem solving) or all six domains (reading, mathematics, science, problem solving, digital reading and digital mathematics).

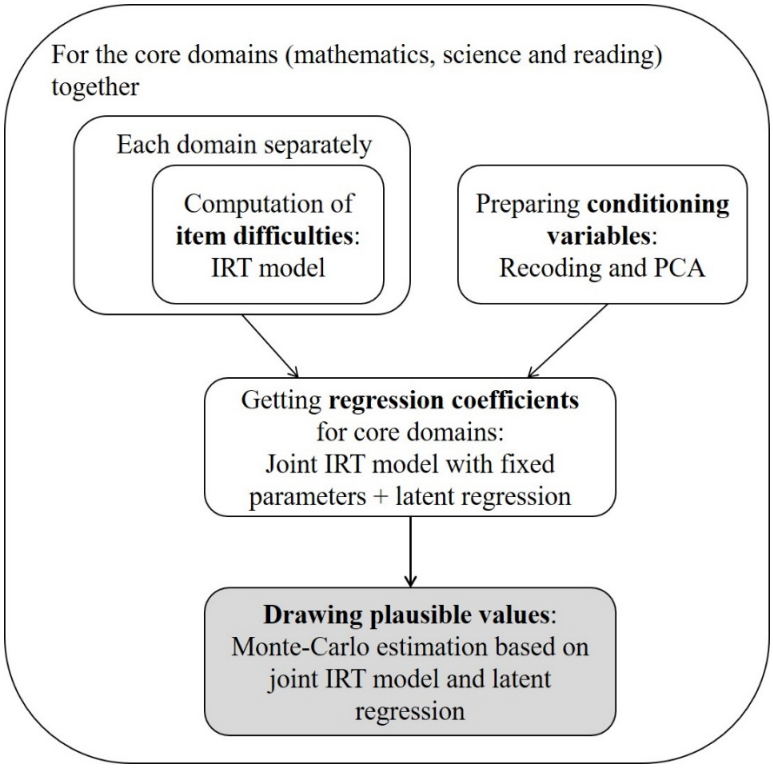


Figure B.1.Computation process of the plausible values, if the country only administered the three core domains

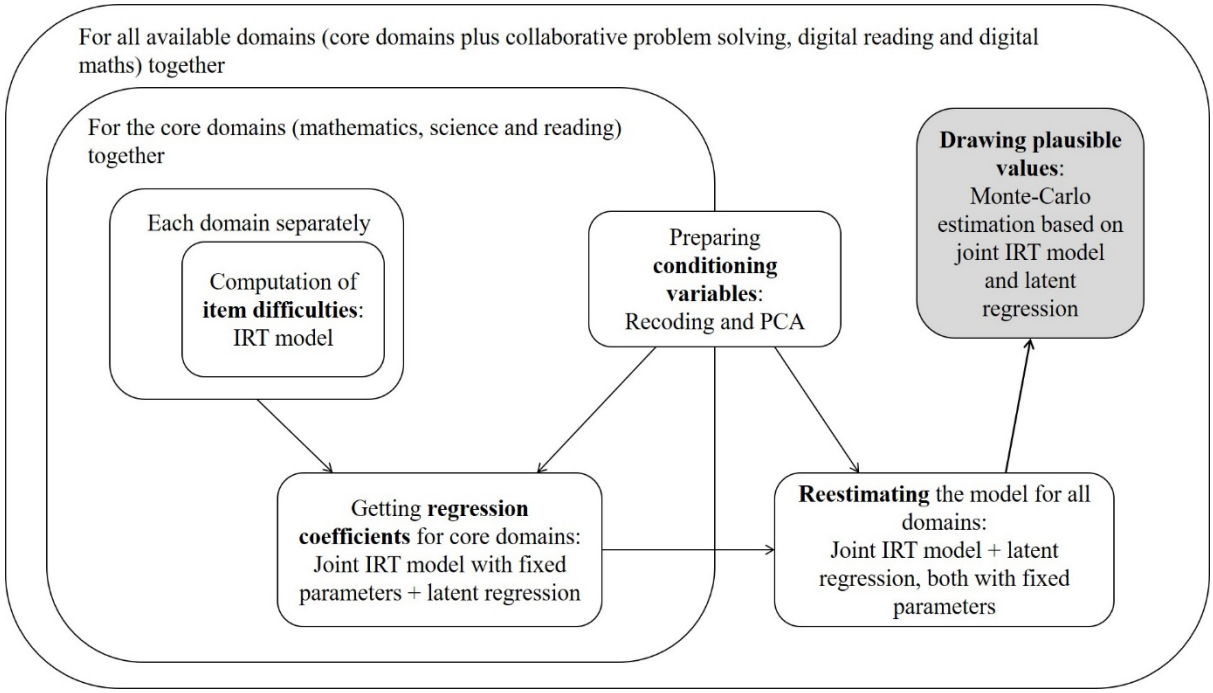


Figure B.2. Computation process of the plausible values, if the country administered additional domains (problem solving and/or digital reading and mathematics) to the three core domains

Appendix C. Computational details of the conducted analysis

This appendix attempts to make the computational procedures we have used as transparent as possible. All code used within our analysis is available from https://github.com/lrzieger/. Our empirical approach used the following steps:

0. Test data preparation. The already scored cognitive data set was downloaded from the OECD homepage (http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm). Subsequent checks were conducted if the data of deleted items was removed (OECD, 2014a, pp. 231, 232) and if missing data was coded correctly (omitted and invalid treated as incorrect and not reached treated as missing; OECD, 2014a, pp. 233, 399).
1. Item difficulty estimation. As this is not the focus of this paper and we do not want the conditioning model to be influenced by estimation of our own item difficulties. We therefore chose to use the published item difficulties within our analysis (Annex A; OECD, 2014a)[14].
2. Preparation of conditioning variables. The conditioning variables were computed from all available questionnaires, for each country and each assessment booklet. We used a two-stage process: recoding (stage 1) and pre-processing (stage 2). For the recoding and first pre-processing, we adhere to the recoding procedures as described in the Annex B in the PISA 2012 technical report (OECD, 2014a, pp. 421–431). The recoding is done for each country separately. The recoded versions of the following variables were used as direct regressors in the later latent regression: Booklet ID, gender, school, grade as well as mother's and father's International Socio-Economic Index (OECD, 2014a, p. 157). The remaining variables were then used within a principal component analysis (PCA) using a singular value decomposition and the correlation matrix. As the technical report does not mention any special adaption of the PCA to account for the categorical nature of some variables, we do not use polychoric correlations. In other words, we try to stay as close to the PISA technical report as possible (OECD, 2014a, p. 157). From this PCA, within each country we retained enough principal components to explain 95% of the variance in the data. This resulted in up to ten principal components being extracted (and a minimum of two) within each country. As this low number of principle components can be surprising for some, see Appendix H 'Number of principle components dependent on student questionnaire booklets' for further information. The conditioning variables are composed of the direct regressors and principal components.
3. Student score estimation. At this point, countries with large samples (over 10,000 students) were split into smaller groups, usually based upon the stratification variables (OECD, 2014a, p. 157). As a consequence, we split the data of those countries into subsamples by alternatingly assigning strata to the new data sets starting with the largest strata.

[14] It is worth noticing that we believe that the step difficulty of item PM155Q03D is a typing error. We substituted the value with the average value across all cycles before where it was used ($\tau_1 = 0.184, \tau_2$=-0.184).

4. The conditioning models are then computed, using a "divide-and-conquer" approach (Patz & Junker, 1999; van Rijn, 2018). This means that we first estimate the IRT model and then estimate the latent regression. This is the default approach used in most large-scale assessments as it is comparatively efficient in terms of computational effort (van Rijn, 2018)[15]. We still experience computational difficulties in four countries (South Korea, Liechtenstein, Columbia and Serbia) leading to missing data for those countries in some of the variations of the conditioning model. The functions `tam.mml()` and `tam.latreg()` from 'TAM' are used to estimate the IRT model and the latent regression. Quasi Monte Carlo integration (Pan & Thompson, 2007) with 2000 nodes and convergence criterions of .001 for deviance and .0001 for the coefficients is used within the computations.
5. Drawing of plausible values. We draw five plausible values for each domain for each student. It is assumed that individual achievement distribution follows a multivariate normal distribution. The distributions are estimated by Monte Carlo estimation with 2000 ability nodes (OECD, 2014a, p. 146).
6. Transformation of plausible values to scale. Again, the transformation of the plausible values to the common PISA scale is not in the focus of this paper. Therefore, we use the formulas from the technical report (OECD, 2014a, pp. 253, 254). For the sake of convenience, we also use the placement on the PISA scales.

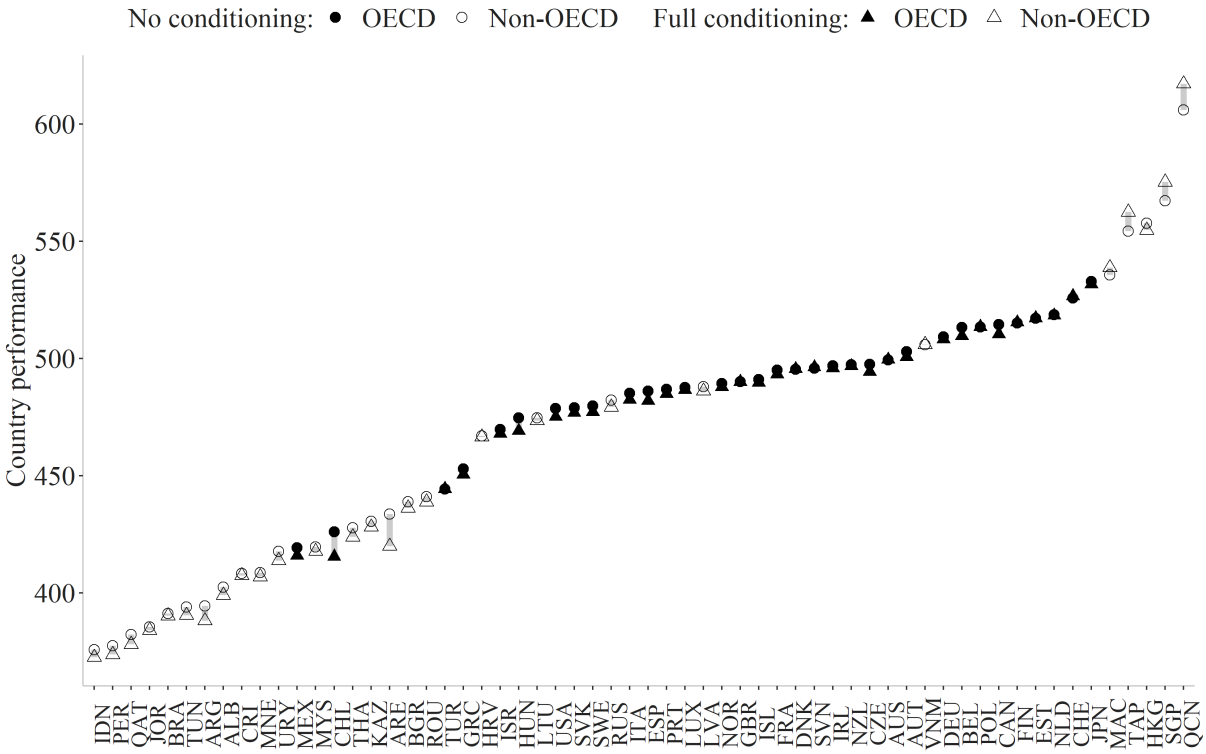
The computations are not deterministic and are therefore influenced by a certain amount of random error (e.g. in randomly drawing plausible values). To make the computations reproducible, we set seeds for the computation. We reran the analysis with different seeds but ended up with similar conclusions.

[15] This approach does have some limitations, however. For instance, it ignores the uncertainty in parameter estimates within the latent regression.

Appendix D. Mathematics: Domain specific analyses

Average scores

Figure D.1 highlights that the country average scores in mathematics are not sensitive to the specification of the conditioning model. The markers for no conditioning (triangle) and full conditioning (circle) sit on top of each other with only few exceptions (e.g. United Arab Emirates and Chile), but even in those cases the difference between the two scores is comparatively small. Hence, the ranking of countries also remains roughly the same in mathematics independent of the conditioning model specification.



Notes: Triangles provide estimates without conditioning and circles with conditioning. Solid markers are OECD countries and hollow markers non-OECD countries.

Figure D.1. Country average mathematics scores with and without conditioning

Table D.1 shows the average mathematics scores when using conditioning models M0-M7 and should be read vertically. The colours depict the scores relative to the other countries' scores with a green (red) value corresponding to a higher (lower) relative score. The colour scheme across the different conditioning model specifications is similar, which means that there is not much change between the specifications. This is confirmed by correlations between 0.99 and 1 between the different specifications. In contrast to reading, the OECD average score also maintains a similar level dropping only 2 points from no conditioning (492 points) to full conditioning (490 points).

Table D.1. Variation in estimated average PISA mathematics scores by conditioning model specification. OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
South Korea	549	557	551	-	550	-	-	-
Japan	533	532	532	530	532	531	532	532
Switzerland	526	527	528	527	527	526	527	527
Netherlands	519	519	519	519	519	519	518	518
Estonia	517	518	521	516	518	521	510	517
Finland	515	515	516	516	515	516	516	515
Canada	514	510	509	513	509	514	507	510
Belgium	513	513	512	509	510	509	512	510
Poland	513	508	513	510	513	510	513	513
Germany	509	509	514	509	508	509	512	508
Austria	503	503	502	501	502	500	501	501
Australia	499	500	497	500	500	498	501	500
Czech Republic	497	495	495	495	494	495	494	494
Ireland	497	501	491	495	495	495	495	496
New Zealand	497	497	497	497	497	497	497	497
Slovenia	496	492	499	495	496	496	493	496
Denmark	495	500	500	495	498	495	496	496
France	495	492	496	494	494	492	491	493
Iceland	491	490	489	491	489	491	490	490
United Kingdom	490	490	490	491	490	491	490	490
Norway	489	490	492	493	492	487	489	488
Luxemburg	488	487	487	487	487	487	486	487
Portugal	487	484	483	482	486	482	483	485
Spain	486	483	484	484	482	483	483	482
Italy	485	484	483	484	482	484	483	482
Sweden	480	477	478	483	478	481	473	477
Slovak Republic	479	475	480	482	480	475	479	477
USA	479	472	472	474	474	474	475	475
Hungary	475	468	469	468	469	468	468	469
Israel	470	468	463	464	465	467	464	468
Greece	453	451	450	451	450	451	451	450
Turkey	444	445	444	443	444	445	443	444
Chile	426	399	418	403	413	408	410	415
Mexico	419	416	416	416	416	416	416	416
OECD average	**492**	**490**	**491**	**488**	**490**	**488**	**488**	**488**
OECD median	**495**	**492**	**494**	**494**	**494**	**492**	**491**	**493**
Correlation with M0	**1.00**	**0.99**	**0.99**	**0.99**	**1.00**	**0.99**	**0.99**	**1.00**
Correlation with M7	**1.00**	**0.99**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**

Notes: Figures illustrate how average PISA mathematics scores vary depending upon the specification of the conditioning models. Green shading indicates higher scores relative to other countries and red cells lower scores. The average mathematics score for non-OECD countries can be found in Table D.2. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. South Korea is missing scores due to computational difficulties.

Table D.2. Variation in estimated average PISA mathematics scores by conditioning model specification. Non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
ALB	402	400	399	400	399	400	399	399
ARE	434	413	415	417	415	415	419	420
ARG	394	389	390	390	388	389	389	388
BGR	439	437	436	437	436	437	436	436
BRA	391	390	389	390	390	391	390	390
COL	387	-	376	372	-	367	372	-
CRI	408	408	407	406	408	407	407	407
HKG	558	563	565	564	563	562	558	555
HRV	467	467	466	466	467	467	466	466
IDN	376	373	372	373	372	373	372	372
JOR	385	384	383	384	384	384	384	384
KAZ	430	430	427	429	428	430	427	428
LIE	530	529	530	531	529	-	529	-
LTU	475	474	473	474	473	474	473	473
LVA	488	487	486	487	486	487	486	486
MAC	536	540	540	539	539	539	539	539
MNE	409	407	406	407	407	408	406	407
MYS	420	419	418	418	418	419	418	418
PER	377	375	374	374	374	375	374	374
QAT	382	380	378	380	378	380	378	378
QCN	606	604	618	619	617	613	619	617
ROU	441	440	439	440	439	440	439	439
RUS	482	477	479	476	478	478	481	479
SGP	567	576	569	576	576	575	577	575
SRB	447	447	446	-	446	-	-	-
TAP	554	564	564	560	563	560	563	562
THA	428	424	424	425	423	425	424	424
TUN	394	392	390	392	390	392	390	390
URY	418	415	414	415	414	415	414	414
VNM	506	506	506	506	506	507	506	506

Notes: Figures illustrate how average PISA mathematic scores vary depending upon the specification of the conditioning models. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. Columbia, Liechtenstein and Serbia are missing the scores due to computational difficulties.

Inequality in PISA scores

We are not only interested in the country average scores, but also in inequality measures. The difference between the 90th and 10th percentile is an inequality measure for spread and displayed in Table D.3. The table vertically depicts the percentile differences according to the different specification. The colours denote lower (higher) inequality in green (red) in relation to the other countries per specification.

While the average mathematics scores are not sensitive to the specification, the percentile gaps are. This becomes obvious through the changes in colours between the columns. The greatest difference exists between no conditioning and conditioning (M1-M7), which is also reflected through rather low correlations roughly between 0.7 and 0.8. Furthermore, the average OECD percentile difference experiences a sharp rise from 214 to a somewhere between 248 and 253 as soon as any form of conditioning is applied.

Even though the major differences are between no conditioning and any form of conditioning, there is also relative changes between the different specifications (see differing colour patterns). The correlation between the conditioning specifications ranges between 0.8 and 1 with especially high correlations ($r >$ 0.9) when direct regressors are included. In line with the results from reading, the direct regressors seem to be an important part of the conditioning model.

Table D.3. Estimates of inequality in PISA mathematics scores across countries by specification of the conditioning model (P90 – P10 gaps). OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	159	184	190	184	186	183	184	184
Estonia	190	248	238	243	241	230	234	235
Chile	193	196	216	198	229	230	210	224
Denmark	195	221	258	252	233	250	246	246
Finland	196	216	214	217	215	217	217	219
Ireland	196	252	228	252	247	252	250	247
Spain	202	224	232	225	234	226	232	233
Greece	203	227	228	227	229	227	229	229
Canada	204	243	242	241	229	234	234	226
Norway	208	268	276	276	276	268	276	276
Sweden	210	174	234	214	232	227	225	232
Slovenia	211	274	274	267	239	271	238	250
Iceland	214	237	240	235	240	236	239	238
USA	214	263	263	268	267	269	269	269
Austria	217	272	281	272	271	267	276	267
Italy	217	241	246	241	247	242	246	246
Poland	217	238	236	274	239	272	239	240
United Kingdom	218	240	244	241	243	241	243	243
Japan	218	233	239	229	238	236	237	238
Switzerland	220	245	244	244	246	246	244	247
Netherlands	220	240	242	241	242	242	242	242
Turkey	220	240	236	243	239	241	242	240
Portugal	221	247	279	283	258	279	282	259
Hungary	222	277	282	276	277	275	279	274
Luxemburg	224	245	248	245	248	245	245	244
France	226	282	284	291	274	278	277	263
Australia	227	279	269	280	247	256	281	250
Czech Republic	228	246	259	246	256	246	258	255
Germany	229	292	294	291	267	286	296	264
New Zealand	233	254	258	255	257	256	258	258

Belgium	241	306	293	304	273	299	296	276
Slovak Republic	242	271	311	283	298	280	309	292
Israel	246	297	284	306	289	295	285	269
OECD average	**215**	**248**	**253**	**253**	**249**	**252**	**252**	**248**
OECD median	**217**	**245**	**246**	**246**	**246**	**246**	**245**	**246**
Correlation with M0	**1.00**	**0.70**	**0.77**	**0.74**	**0.80**	**0.77**	**0.80**	**0.79**
Correlation with M7	**0.79**	**0.81**	**0.93**	**0.86**	**0.97**	**0.90**	**0.94**	**1.00**

Notes: Figures illustrate how the difference between the 90th and 10th percentile of PISA mathematics scores changes depending upon the specification of the conditioning model. The mathematics percentile differences for non-OECD countries can be found in Table D.4. Green shading indicates less inequality in reading scores relative to other countries and red cells greater inequality. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Table D.4. Estimates of inequality in PISA mathematics scores across countries by specification of the conditioning model (P90 – P10 gaps). Non-OECD countries.

Country	M0	M1	M2	M3	M4	M5	M6	M7
ALB	183	211	211	211	211	211	210	211
ARE	209	254	244	258	247	259	256	255
ARG	165	189	189	189	190	190	190	190
BGR	222	243	242	244	242	243	243	241
BRA	176	197	199	198	195	196	197	194
CRI	145	172	170	171	170	172	171	172
HKG	218	285	282	284	285	286	246	247
HRV	207	228	221	230	224	231	229	230
IDN	152	179	180	179	178	180	180	178
JOR	169	193	194	195	192	193	194	192
KAZ	156	180	181	182	180	180	182	180
LTU	210	233	233	233	233	233	233	232
LVA	188	212	215	213	213	212	215	214
MAC	217	258	263	267	264	266	270	268
MNE	186	209	210	208	209	210	207	209
MYS	186	207	210	208	209	208	209	209
PER	181	208	211	209	209	208	210	209
QAT	229	251	259	250	257	250	255	254
QCN	239	309	316	310	310	305	316	310
ROU	188	210	209	211	208	211	209	209
RUS	202	218	226	221	217	225	216	221
SGP	247	316	271	318	318	320	320	319
TAP	275	347	356	349	352	350	360	357
THA	194	208	216	210	209	208	211	207
TUN	174	197	202	199	196	196	200	196
URY	199	223	224	223	223	223	223	222
VNM	196	222	222	220	223	222	221	221

Notes: Figures illustrate how the difference between the 90th and 10th percentile of PISA mathematic scores changes depending upon the specification of the conditioning model. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

The association between PISA scores and background characteristics

One of the key motivations for using a conditioning model is to counteract attenuation in group estimates. In this paper, we examine if conditioning has an influence on the gaps in gender and migrant status. Gender is a direct regressor while migrant status is processed into indirect regressors (by using principle component analysis).

Figure D.2 highlights how the country gender gaps (regression of mathematics performance upon an indicator of whether the student is female) are influenced by the specification of the conditioning model. Almost all countries experience a negative shift as soon as conditioning is used. Without conditioning no gender differences can found (0 points on average), while boys perform 5 to 6 points better than girls when conditioning with the individual direct regressors included (M2, M4, M6 and M7) is used. It is interesting that nearly all countries experience a negative shift, even when the gender gap from M0 is positive. This means that, despite conditioning attenuation, is still present in some cases, e.g. Finland has a gender difference of 9 points without conditioning, but only a gender gap of 2 points with full conditioning. Overall, the diamonds (M2) and triangles (M7) mostly sit on top of each other and are distinct from the circles (M0) meaning that the gender gap in most countries is not sensitive to exact specification of the model as long as direct regressors are included.

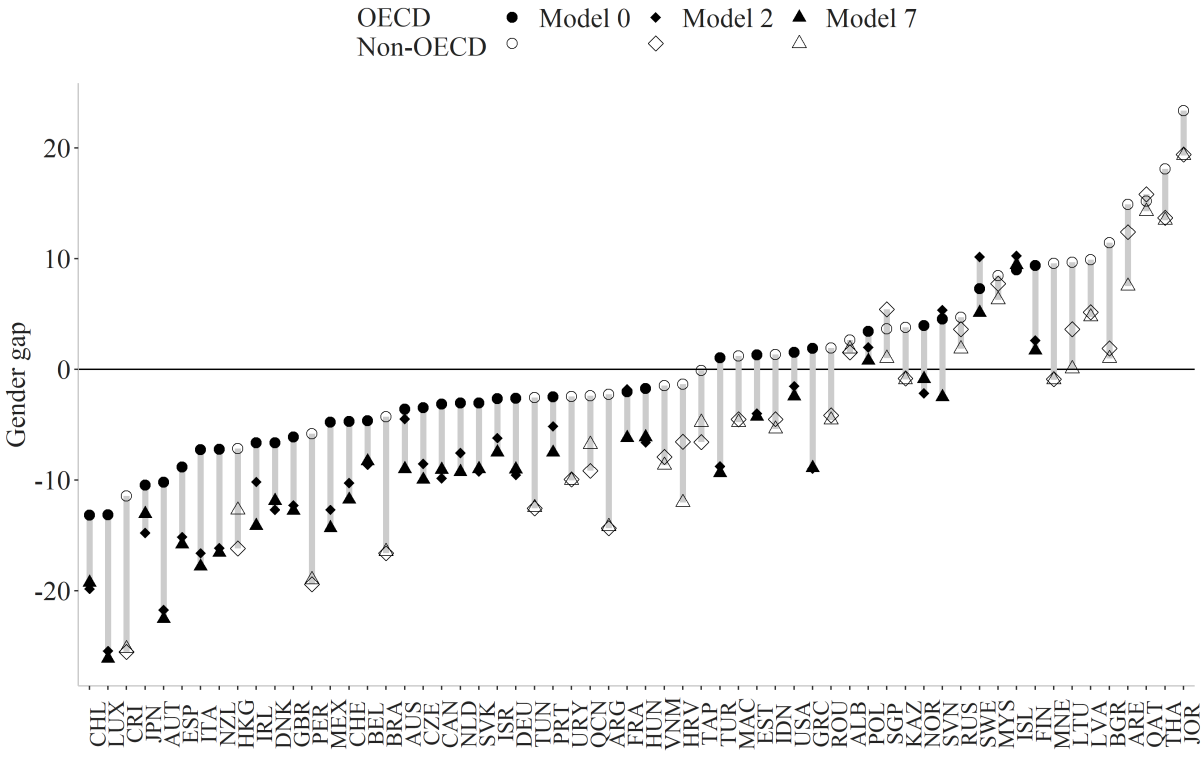


Figure D.2. Country mathematics gender gap without conditioning (M0), just with individual direct regressor including gender (M2) and with full conditioning (M7)

The achievement difference between migrant and native students in mathematics can be seen in Figure D.3. Overall, the three symbol – circle (M0), diamond (M3) and triangle (M7) – sit roughly on top of each other signalling that there are relatively small changes in the migrant-native gap between the specifications. The average migrant-native gap drops from -20 points (M0) to -24 (M3) and -24 points (M7), but the gaps in the countries itself cover a rather big range from -130 points (M7 in Shanghai) to +89 points (M7 in Qatar). One could assume that the migrant-native gap in the mathematics is also not sensitive to the specification of the conditioning model.

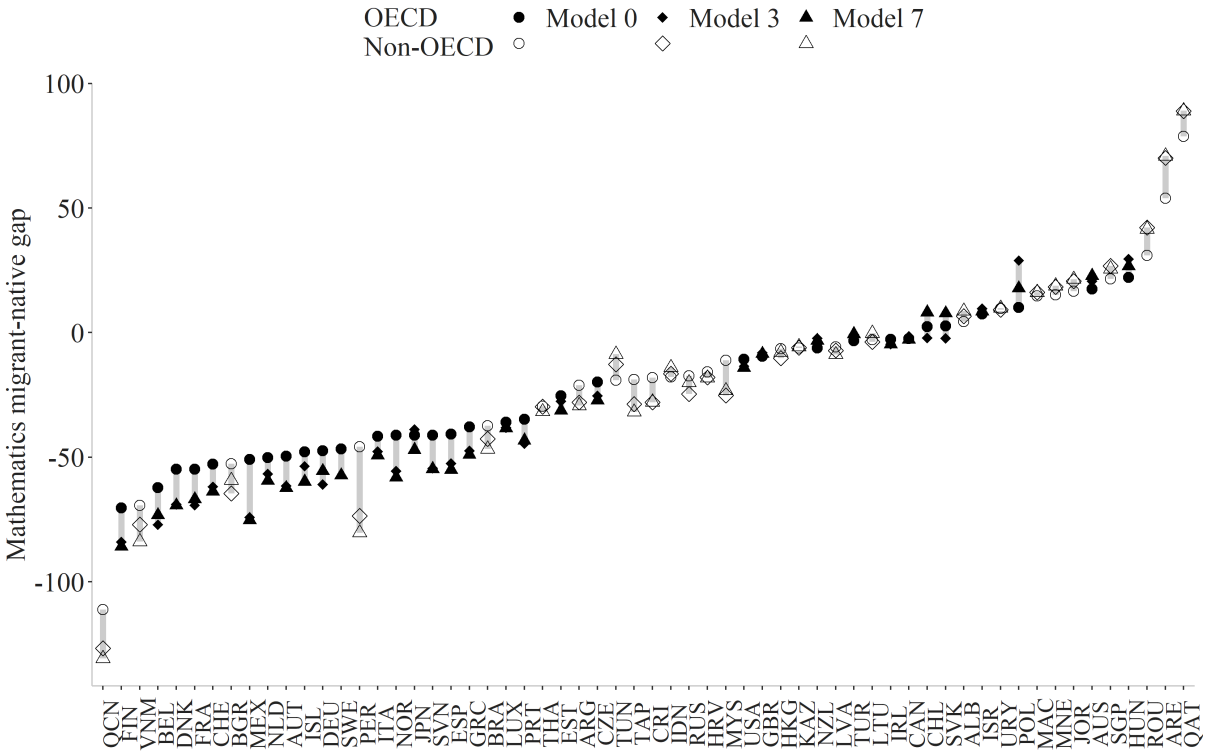


Figure D.3. Country mathematics gap between migrant and native students without conditioning (M0), with indirect regressors (migration status was pre-processed) in conditioning (M3) and with full conditioning (M7)

Appendix E. Science: Domain specific analyses

Average scores

Figure E.1 depicts the country science average scores without conditioning (triangle) and with full conditioning (circle) as well as its difference (line in between). While the conditioning model has more impact on the average scores in science than in mathematics, differences are fairly minor. On average, the scores rise by two points when full conditioning is applied, but there is no common direction. At the extremes, Russia experiences an increase of 12 points, while Tunisia experiences a decrease of -11 points. Yet the ranking of countries remains roughly the same.

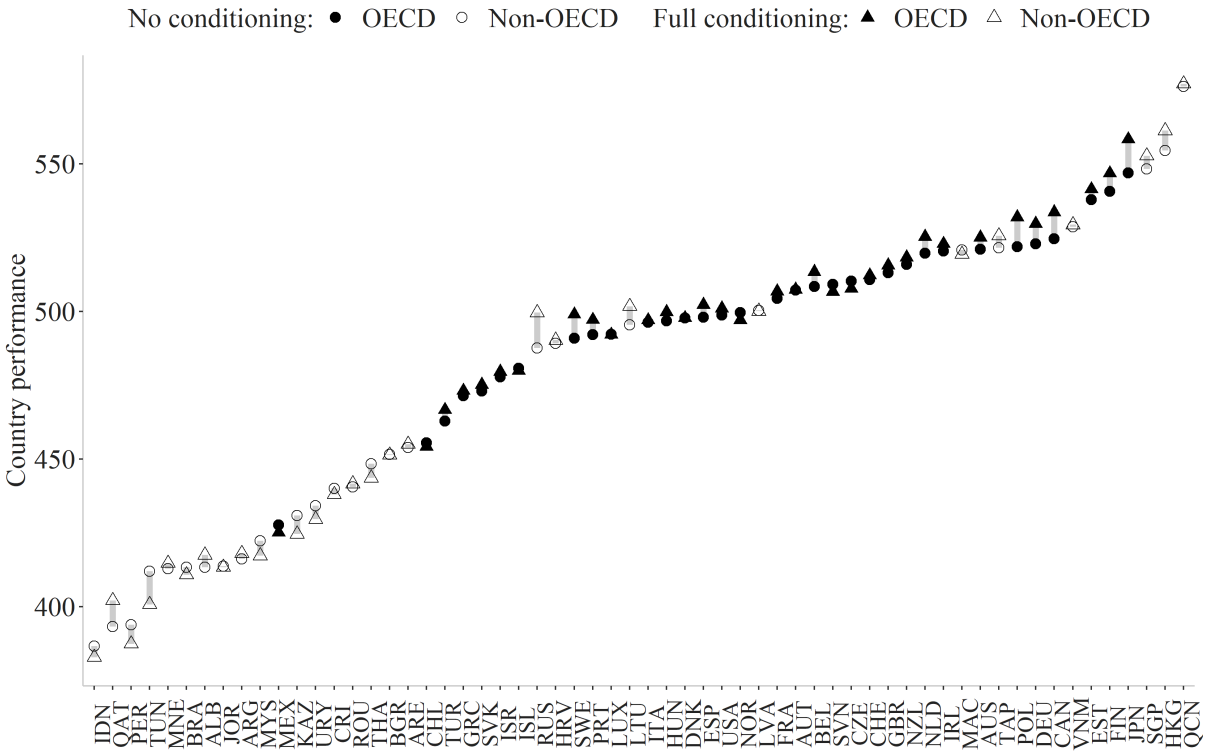


Figure E.1. Country average science scores with and without conditioning

This is stressed by Table E.1, which shows a rather consistent colour scheme and only minor variation in relative scores of the OECD countries. The table should be read vertically inside the conditioning model specification with green (red) scores belonging to higher (lower) relative country average scores. The correlations between all specifications (including no conditioning) is 0.97 or higher. While there is some change, the scores stay reasonably similar across all specifications. The OECD average rises by 2 points from no conditioning (502 points) to full conditioning (504 points).

Table E.1. Variation in estimated average PISA science scores by conditioning model specification. OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Japan	547	548	546	550	560	549	553	558
Finland	541	545	542	544	540	545	546	547
Estonia	538	538	543	533	541	534	546	541
South Korea	535	533	545	-	546	-	-	-
Canada	525	527	530	517	534	523	533	534
Germany	523	522	525	521	531	522	523	530
Poland	522	531	535	526	531	525	531	532
Australia	521	524	523	522	526	527	520	525
Ireland	520	521	528	520	524	520	524	523
Netherlands	520	520	513	520	524	520	519	525
New Zealand	516	516	515	516	516	516	518	518
United Kingdom	513	515	508	514	512	515	514	516
Switzerland	511	515	512	513	510	514	510	512
Czech Republic	510	507	501	509	509	507	509	508
Slovenia	509	509	490	508	520	506	519	507
Belgium	508	507	507	509	514	510	514	513
Austria	507	509	501	507	511	506	507	507
France	504	502	505	510	503	503	511	507
Norway	500	499	499	500	498	500	498	497
USA	499	505	503	499	502	497	501	501
Denmark	498	496	488	488	497	495	497	498
Spain	498	498	506	497	505	497	504	502
Hungary	497	491	495	493	499	491	498	500
Italy	496	494	498	495	498	494	496	497
Luxemburg	492	492	492	492	492	492	492	492
Portugal	492	493	497	485	499	485	492	497
Sweden	491	509	503	496	501	495	503	499
Iceland	481	480	479	480	479	480	479	480
Israel	478	470	485	472	489	472	488	480
Slovak Republic	473	476	476	466	476	473	473	475
Greece	471	470	474	471	473	470	473	473
Turkey	463	462	462	463	460	462	461	467
Chile	455	449	458	452	453	447	449	454
Mexico	428	426	422	425	424	426	424	425
OECD average	**502**	**503**	**503**	**500**	**506**	**501**	**504**	**504**
OECD median	**506**	**507**	**503**	**507**	**507**	**503**	**507**	**507**
Correlation with M0	**1.00**	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Correlation with M7	**0.99**	**0.99**	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**	**1.00**

Notes: Figures illustrate how average PISA science scores vary depending upon the specification of the conditioning models. The average science score for non-OECD countries can be found in Table E.2. Green shading indicates higher scores relative to other countries and red cells lower scores. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. South Korea is missing scores due to computational difficulties.

Table E.2. Variation in estimated average PISA science scores by conditioning model specification. Non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
ALB	413	412	417	411	417	412	417	417
ARE	454	458	458	457	458	457	455	455
ARG	416	414	415	412	408	413	420	418
BGR	452	450	454	450	447	450	454	451
BRA	413	414	404	412	410	413	406	411
COL	419	-	411	419	-	414	409	-
CRI	440	439	434	439	436	438	435	438
HKG	554	556	557	556	557	555	556	561
HRV	489	488	497	489	493	488	489	490
IDN	387	385	377	384	385	384	377	383
JOR	414	414	414	412	410	413	410	413
KAZ	431	428	424	429	419	428	434	425
LIE	526	528	523	534	536	-	531	-
LTU	495	495	501	495	502	495	497	502
LVA	500	500	499	500	499	500	499	500
MAC	521	520	522	520	520	519	520	519
MNE	413	412	411	411	417	411	418	415
MYS	422	420	414	419	418	420	415	417
PER	394	391	385	391	389	390	385	387
QAT	393	391	403	391	400	391	400	402
QCN	576	579	580	574	578	571	580	577
ROU	441	440	437	440	443	439	438	442
RUS	488	499	498	505	503	498	497	500
SGP	548	555	534	552	552	556	551	553
SRB	449	448	450	-	448	-	-	-
TAP	522	527	527	518	525	521	525	526
THA	448	446	445	445	444	445	444	444
TUN	412	410	409	410	400	409	406	401
URY	434	432	438	431	443	431	437	430
VNM	529	529	528	529	528	528	530	529

Notes: Figures illustrate how average PISA science scores vary depending upon the specification of the conditioning models. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. Columbia, Liechtenstein and Serbia are missing the scores due to computational difficulties.

Inequality in PISA scores

In contrast to the country average scores, the percentile gap (P90-P10) experiences substantial changes depending on the specification of the conditioning model. This becomes clear when assessing table E.3, which again depicts relative scores with green (red) scores relating to lower (higher) inequality. The mixed colouring and big changes between the columns make it apparent that the scores and the countries

relative positions change substantially depending on the used specification. Countries, such as the Slovak Republic, which was in the bottom category (high inequality) for some specifications (M0, M1 and M5) end up in the top category for others (M2, M4 and M6).

Table E.3. Estimates of inequality in PISA science scores across countries by specification of the conditioning model (P90 – P10 gaps). OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	140	164	157	163	160	165	162	163
Estonia	181	104	96	126	110	124	138	133
Chile	184	146	160	146	174	168	184	176
Turkey	185	208	205	206	207	208	205	210
Spain	188	211	203	212	204	213	207	208
Greece	196	224	215	225	221	225	224	228
Poland	197	179	209	176	218	180	217	214
Canada	201	151	137	147	171	165	173	178
Switzerland	205	229	214	229	219	230	218	220
Czech Republic	206	225	217	224	215	228	216	218
Hungary	206	189	200	191	207	191	206	209
Ireland	207	147	176	164	187	169	186	192
Portugal	207	196	216	193	209	204	189	212
Italy	209	235	223	236	225	237	226	230
Denmark	210	125	148	162	139	173	167	177
Austria	212	166	185	194	196	206	191	205
Slovenia	212	202	156	177	209	205	199	222
Finland	214	234	235	231	233	230	228	228
USA	214	183	186	170	185	183	181	186
Japan	215	229	228	228	260	231	254	260
Sweden	218	306	248	237	239	233	198	217
France	219	196	180	172	221	209	218	231
Germany	220	175	145	175	213	206	162	224
Norway	220	163	169	170	170	188	179	185
Iceland	223	248	223	252	226	252	226	229
Netherlands	223	245	226	243	233	245	230	236
Belgium	226	181	256	193	206	201	262	227
United Kingdom	228	253	235	252	240	255	242	248
Australia	232	204	211	206	231	226	216	238
Slovak Republic	235	222	171	204	196	239	179	216
Luxemburg	236	261	253	264	253	265	263	265
New Zealand	238	263	250	263	251	264	254	257
Israel	239	235	280	229	260	236	262	261
OECD average	**210**	**203**	**200**	**202**	**209**	**211**	**208**	**215**
OECD median	**212**	**204**	**209**	**204**	**213**	**209**	**207**	**218**
Correlation with M0	**1.00**	**0.49**	**0.51**	**0.52**	**0.59**	**0.65**	**0.54**	**0.70**
Correlation with M7	**0.70**	**0.79**	**0.82**	**0.83**	**0.96**	**0.89**	**0.88**	**1.00**

Notes: Figures illustrate how the difference between the 90th and 10th percentile of PISA science scores changes depending upon the specification of the conditioning model. The science percentile gaps for non-OECD countries can be found in Table E.4. Green shading indicates less inequality in reading

scores relative to other countries and red cells greater inequality. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Table E.4. Estimates of inequality in PISA science scores across countries by specification of the conditioning model (P90 – P10 gaps). Non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
ALB	188	220	225	219	225	220	223	223
ARE	210	177	175	198	177	188	204	206
ARG	177	206	207	206	210	207	205	209
BGR	237	262	261	260	262	261	259	260
BRA	163	187	190	187	188	188	189	189
CRI	139	164	165	164	166	166	165	167
HKG	180	108	94	99	113	118	176	175
HRV	197	219	213	220	213	221	217	220
IDN	140	166	167	165	166	168	166	166
JOR	177	201	202	205	201	204	203	203
KAZ	156	188	188	187	190	188	184	186
LTU	196	220	216	221	208	221	215	223
LVA	176	199	187	196	190	198	191	193
MAC	177	105	101	113	110	121	112	121
MNE	189	213	222	214	219	214	219	219
MYS	175	200	199	202	202	203	203	203
PER	153	179	176	179	179	181	179	181
QAT	239	264	265	265	264	266	257	266
QCN	188	152	113	140	145	174	118	148
ROU	177	202	202	203	200	203	203	201
RUS	189	180	172	174	192	188	167	200
SGP	239	171	236	171	171	181	170	177
TAP	196	147	134	160	139	160	136	140
THA	171	190	191	193	190	191	193	192
TUN	169	196	207	196	205	196	201	204
URY	199	225	234	229	230	231	229	235
VNM	171	194	184	194	187	195	190	193

Notes: Figures illustrate how the difference between the 90th and 10th percentile of PISA science scores changes depending upon the specification of the conditioning model. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

The association between PISA scores and background characteristics

One of the key motivations for using a conditioning model is to counteract attenuation in group estimates. In this paper, we examine if conditioning has an influence on the gaps in gender and migrant status. Gender is a direct regressor while migrant status is processed into indirect regressors (by using principle component analysis).

Figure E.2 shows that the conditioning model specifications have rather little influence on the gender gap in science, especially in comparison to the gender gaps in mathematics and reading. In multiple countries the three symbols sit on top of each other or close together which means that the gender gap is robust against the specification. For the remaining countries, (substantial) change can be seen depending on the specification, but there is no common direction or magnitude of the science gender gap. Overall, the average gender gap rises only one point between no conditioning (3 points) and full conditioning (4 points). Yet, because not all countries are sensitive to the specification of the conditioning model, distinct changes in rank can occur.

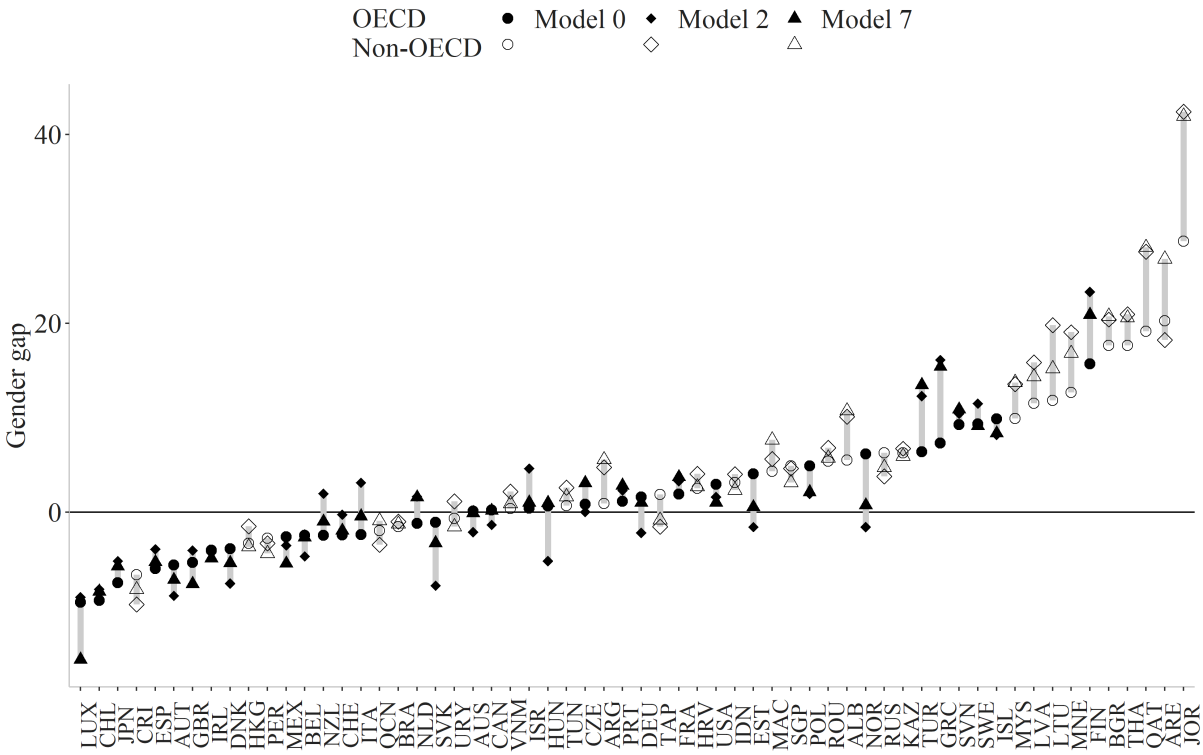


Figure E.2. Country science gender gap without conditioning (M0), just with individual direct regressor (incl. gender) in conditioning (M2) and with full conditioning (M7)

Figure E.3 displays the gaps in science achievement for another grouping variable – migrant status (native vs migrant students). Again, the influence of the conditioning model specification is rather small with even more countries, which have the three symbols sitting on top of each other or close together.

In the countries, which exhibit a wider spread of symbols, it is usually the diamond (M3; just indirect regressors) which is further apart with the circle (M0) and the triangle (M7) staying closer together.

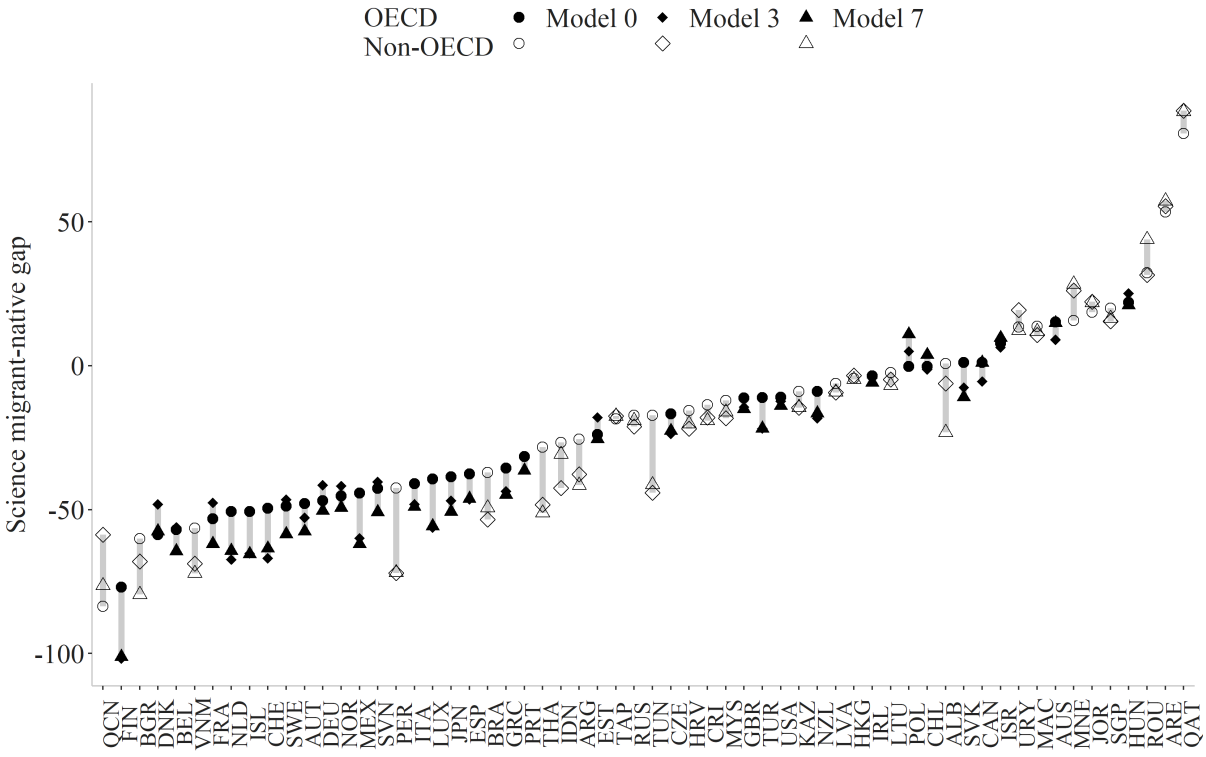


Figure E.3. Country science gap between migrant and native students without conditioning (M0), with indirect regressors (migration status was pre-processed) in conditioning (M3) and with full conditioning (M7)

Appendix F. Reading: Non-OECD specific tables

Table F.1. Variation in estimated average PISA reading scores by conditioning model specification. Non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
ALB	407	404	419	405	414	404	417	397
ARE	444	476	469	469	475	474	467	463
ARG	405	401	419	401	414	401	403	407
BGR	440	437	441	437	428	437	435	448
BRA	414	415	420	413	418	414	421	419
COL	421	-	447	452	-	448	455	-
CRI	448	446	453	446	447	445	449	447
HKG	542	535	525	531	524	537	506	528
HRV	483	483	445	483	449	482	459	470
IDN	399	397	407	397	395	396	406	393
JOR	404	401	414	401	411	401	414	407
KAZ	395	392	401	392	395	392	396	388
LIE	512	511	479	502	481	-	493	-
LTU	475	475	446	474	441	474	452	471
LVA	489	487	459	487	462	486	465	466
MAC	509	493	488	492	488	493	488	490
MNE	425	424	435	422	434	423	428	428
MYS	403	400	410	399	406	399	408	402
PER	401	398	397	397	398	397	403	396
QAT	398	395	378	395	375	395	375	376
QCN	565	588	561	581	549	575	557	548
ROU	435	434	443	434	442	434	442	436
RUS	477	476	459	473	452	476	460	450
SGP	537	531	499	529	518	534	514	516
SRB	450	449	438	-	460	-	-	-
TAP	520	512	507	527	506	522	504	504
THA	443	440	434	441	438	440	437	440
TUN	414	412	428	412	427	412	427	427
URY	426	424	428	423	426	423	425	426
VNM	505	505	488	505	487	505	490	497

Notes: Figures illustrate how average PISA reading scores vary depending upon the specification of the conditioning models. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. Columbia, Liechtenstein and Serbia are missing the scores due to computational difficulties.

Table F.2. Estimates of inequality in PISA reading scores across countries by specification of the conditioning model (P90 – P10 gaps). Non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
ALB	216	256	253	254	256	256	254	277
ARE	215	175	195	189	179	180	198	204
ARG	201	232	240	237	238	237	246	236
BGR	272	308	314	310	324	310	316	314
BRA	188	219	218	218	218	219	219	219
CRI	153	183	182	183	182	182	180	184
HKG	192	203	202	205	207	205	201	231
HRV	194	220	229	219	230	219	226	226
IDN	156	190	186	189	190	192	189	191
JOR	198	231	227	230	228	230	230	229
KAZ	158	187	187	187	189	190	189	199
LTU	198	223	227	226	235	225	232	229
LVA	184	211	218	211	219	211	218	217
MAC	182	212	212	215	214	214	216	215
MNE	204	236	242	234	241	235	238	239
MYS	184	213	213	215	214	215	217	218
PER	197	229	228	233	229	231	233	232
QAT	250	284	276	281	278	284	283	279
QCN	184	176	165	104	176	162	168	175
ROU	203	233	233	233	232	235	239	245
RUS	204	272	271	282	269	273	269	261
SGP	230	219	213	212	218	239	211	226
TAP	214	182	181	147	178	179	184	184
THA	174	196	205	202	200	200	205	201
TUN	192	228	230	228	232	229	231	234
URY	207	241	250	239	250	241	248	248
VNM	162	182	196	183	188	183	194	188

Notes: Figures illustrate how the difference between the 90^{th} and 10^{th} percentile of PISA reading scores changes depending upon the specification of the conditioning model. M0 = no conditioning; M1-M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressor, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Appendix G. How does the migrant-native gap in reading scores change when migrant status is used as a direct (rather than indirect) regressor?

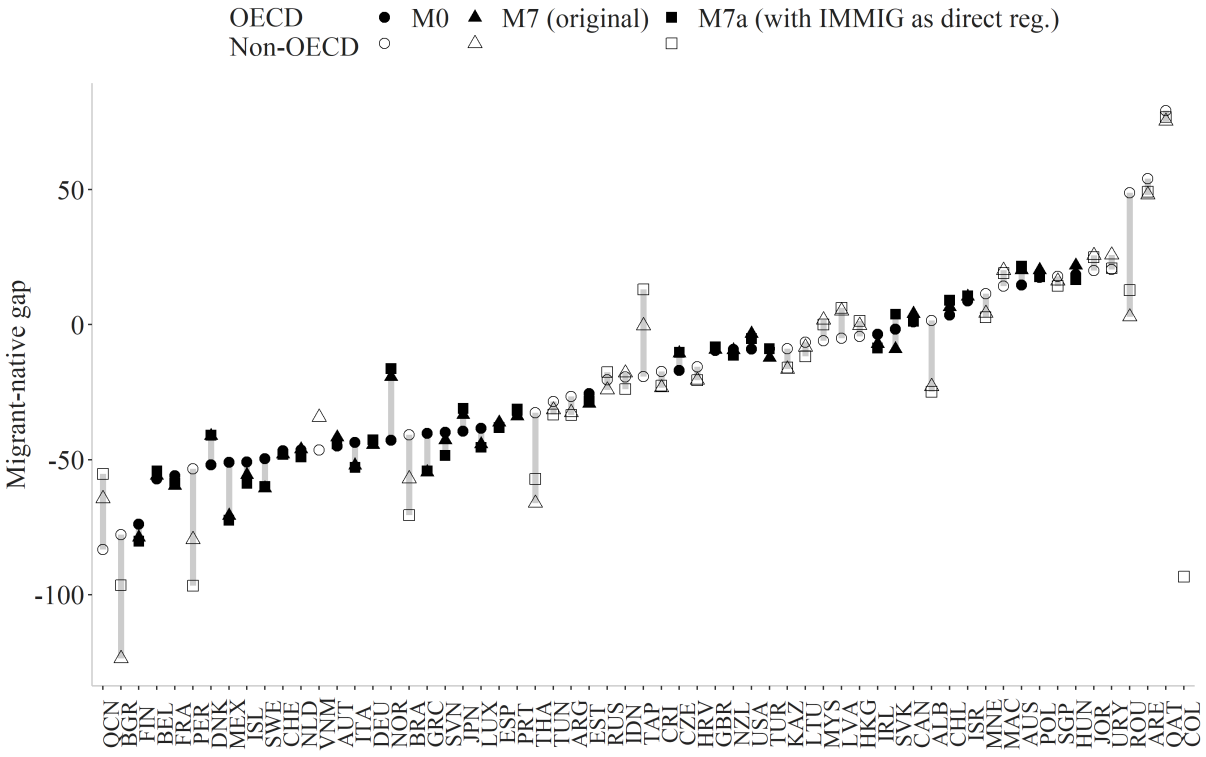
Figure G.1 highlights the differences in the migrant-native gap in reading scores between three separate models:

(i) No conditioning (M0)

(ii) Full conditioning with migrant status as an indirect regressor (M7)

(iii) Full conditioning with migrant status as a direct regressor (M7a)

In most countries the estimated migrant-native gap does not change whether migrant status is used as a direct or indirect regressor (triangle and square on top of each other). Again, however, there are some important individual exceptions. In some countries, such as Bulgaria (M7 = -124; M7a = -96) and Peru (M7 = -80; M7a = -97), there is an appreciable change in at least the magnitude of the immigrant-native gap. These are, however, the exceptions rather than the rule. Overall, it seems that the decision of whether to include immigrant status as a direct or indirect regressor has a trivial impact upon the substantive results.



Notes: Altered model 7a (IMMIG included in direct regressors) could not be computed for Korea due to computational difficulties, because of the very small subset of non-native students.

Figure G.1. Country reading gap between migrant and native students without conditioning (model 0), with original model 7 and altered model 7a (IMMIG included as direct regressor).

Appendix H. Number of principal components dependent on the student questionnaire booklets

In PISA 2012, the rotated design is not only used for the cognitive items but also for the student background questionnaire. Overall, three different versions of the student background questionnaire (booklet A, B and C) were administered (questionnaire booklets can be downloaded from http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm). All three included the common parts about the student (section A) and the student's family and home (section B). Furthermore, all three booklets included questions about learning mathematics (section C), but the booklets differed in extent. Booklet A administered all 21 items about learning mathematics, while booklet B (9 items) and booklet C (14 items) contained different subsets. Booklet A also asked questions about the student's problem-solving experience (section F). Booklet B additionally contained items about the student's mathematics experience (section D), the school (section E) and also problem-solving experience (section F). Booklet C also covered the student's mathematics experience (section D) and school (section E). Roughly a third of each country takes each booklet.

Due to the rotated design, the student background questionnaire experiences a substantial amount of missing data while being the foundation for the indirect regressors in the conditioning model. We are therefore interested in how the number of indirect regressors changes if we look at the separate booklet questionnaires only and not complete rotated design (see Table H.1). Overall, the number of principal components vary between countries and booklets. The maximal amount of principal components was 157 in Italy for booklet B and the minimal was 51 in Liechtenstein for Booklet A. The number of principal components varies between the sample with all booklets and the different subsamples for each booklet, but the numbers lay in a plausible range with no surprising outliers anywhere.

Table H.1. Number of principal components used for conditioning, when using the complete background questionnaire as base or the student questionnaire booklet separately (reduced sample size)

Country	All	Booklet A	Booklet B	Booklet C
AUS	103	91	105	102
AUT	115	104	122	116
BEL	145	137	148	144
CAN	102	92	103	100
CHE	109	100	110	103
CHL	130	117	132	131
CZE	95	85	100	91
DEU	110	102	112	124
DNK	120	111	122	120
ESP	106	94	108	101
EST	94	82	95	93
FIN	113	106	118	114
FRA	82	73	86	85
GBR	84	76	85	81

GRC	102	93	110	103
HUN	133	124	139	137
IRL	115	103	118	113
ISL	78	80	87	89
ISR	85	73	87	91
ITA	153	146	157	149
JPN	80	69	82	85
KOR	142	131	146	144
LUX	116	109	117	114
MEX	136	130	140	133
NLD	90	83	93	92
NOR	90	82	90	92
NZL	97	94	101	100
POL	91	84	100	90
PRT	150	142	152	155
SVK	120	113	129	123
SVN	116	108	119	118
SWE	94	88	97	97
TUR	100	92	104	100
USA	77	69	82	78
Non-OECD countries:				
ALB	69	60	78	81
ARE	86	78	90	87
ARG	95	88	100	99
BGR	84	76	85	85
BRA	89	78	90	89
COL	89	76	91	94
CRI	102	94	106	102
HKG	135	122	136	143
HRV	137	129	139	134
IDN	90	84	92	88
JOR	102	97	108	104
KAZ	84	77	88	80
LIE	55	51	53	82
LTU	81	73	86	79
LVA	121	106	120	119
MAC	139	132	146	142
MNE	83	76	87	85
MYS	82	72	84	82
PER	88	79	91	92
QAT	83	72	85	88
QCN	95	84	98	97
QRS	97	83	93	99
ROU	86	78	91	84
RUS	102	90	104	99
SGP	110	100	116	112
SRB	118	110	118	120
TAP	96	88	102	95
THA	86	77	88	80

TUN	83	76	87	88
URY	109	99	110	111
VNM	81	76	89	83

 @cepeo_ucl

ucl.ac.uk/ioe/cepeo