Essays on the psychometric and statistical properties of the Programme for International Student Assessment

Laura Raffaella Zieger

PhD Social Statistics

Social Research Institute

University College London (UCL)

July 2021

Declaration

I, Laura Raffaella Zieger, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

- Chapters 1 and 5 are single-authored by Laura Raffaella Zieger.
- Chapter 2 is joint work by Laura Raffaella Zieger, John Jerrim, Jake Anders, and Nikki Shure. Laura Raffaella Zieger is the main author, who conceptualised the idea and who undertook all the statistical analyses. John Jerrim, Jake Anders, and Nikki Shure contributed to the chapter through supervision, guidance, and reviews.
- Chapters 3 and 4 are joint work by Laura Raffaella Zieger and John Jerrim. Laura Raffaella Zieger is the main author, who conceptualised the idea and who undertook all the statistical analyses. John Jerrim contributed to the chapters through guidance and reviews.
- 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.

Acknowledgements

I have been very lucky that my PhD project is part of the European Training Network OCCAM. As a result, I benefitted from generous funding, a phenomenal group of peers, interesting trainings and great support and feedback. This journey would not have been the same and only half as good without OCCAM and the fellow PhD students and seniors. Special thanks to Silvan and Andrés who kept me sane during the hard times and made the rest even better.

A very special and deep thanks goes to my current supervisor John Jerrim and my previous boss and mentor Rolf Strietholt for their never-ending support, interesting discussions, and food for thought as well as inspiring me to grow both as an academic and as a person. Furthermore, many thanks to my co-supervisors Jake Anders and Nikki Shure for all the discussions, ideas, and comments.

Finally, I cannot thank my family and my partner, Alex, enough. Without them, I would not be where I am today. Their unconditional love and support mean everything to me and enable me to reach for the stars.

Abstract

The Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) has become one of the key studies for evidence-based education policymaking across the globe. Yet its psychometric and statistical properties – and therefore its limitations and implications – are often poorly understood by researchers and stakeholders. This thesis aims to shed new light on the impact that psychometric and statistical issues can have on key statistics and cross-country comparisons.

In the first study, I investigate the 'conditioning model', where background variables are used in the derivation of student achievement scores. Thereby, I systematically vary the background variables used within the conditioning model and analyse the impact upon country rankings. My key finding is that the exact specification of the conditioning model matters; cross-country comparisons of PISA scores can change substantially depending upon the statistical methodology used.

The second study is an extension of the first, as I conduct a simulation study to further investigate research questions and additionally examine other key aspects influencing the achievement score computation, more specifically the preparation of variables and the test and background questionnaire design. This study confirms that the specification of the conditioning model matters, though bias in the results can be reduced by asking students questions in all subjects. In contrast, variable preparation and the background questionnaire design have negligible impact.

The third study aims to provide a broader and more comprehensive review of different psychometric and statistical properties in PISA. By using a case study, six different properties that can potentially affect the validity of parental education group comparisons in Germany are evaluated. This study highlights how diverse sources of bias can be in PISA, and that they indeed impact the results. While I did not find that all the investigated issues introduce bias, several properties substantially impacted the results.

Impact statement

This thesis is part of the European Training Network Outcomes and Causal Inference in International Comparative Assessments (OCCAM), which conducts international, interdisciplinary and intersectoral research on international comparative education. My research contributes to the work on the integrity of educational measures. Thereby, I investigate different psychometric and statistical properties of the Programme for International Student Assessment (PISA), which can impact the validity and comparability of PISA scores. The findings are hence relevant for three different audiences.

Academia: The three conducted studies address different gaps in the literature. The first study generated insights about the impact that background variables can have upon student achievement scores in PISA through the computation. It was shown that the exact model specification is important, especially in the minor domains. The second study employed a simulation to further investigate questions about the measurement model and related properties, which would not be possible with real-life data. Apart from the confirmation that the model specification matters, one key finding was that bias decreases when students are administered questions in all domains. In contrast, the background questionnaire design and variable preparation had little impact. The third study examined the process of measuring student achievement in a more comprehensive way drawing upon the total survey error framework, which has been seldomly done before for PISA. Six potentially problematic properties were identified and examined in a case study. While not all turned out problematic, some introduced substantial bias into group comparisons.

Users of PISA data: PISA data is used in varied ways, such as in applied research, the media, education debates or as a tool of soft governance. As a result, PISA scores and analyses influence education policies, daily school life and the public opinion on education. To put it in a nutshell, this thesis aimed to illustrate why the PISA data should be used and interpreted (more) carefully. Different psychometric and statistical properties that can bias the results are investigated. Especially the third study, a case study, highlights

how the comparability and validity of commonly used measures can be severely impacted through multiple pathways. Limitations and implications need to be better communicated and heeded, so that appropriate analyses are conducted and informed decisions can be made.

Study makers: Throughout this thesis, two themes emerged which are relevant for study makers of international large-scale assessments. The first is a request for better and more transparent communication of psychometric and statistical properties and their implications. In all three studies, issues relating to the replicability and official communication of properties surfaced. The current practice significantly exacerbates in-depth scrutiny and makes it difficult for non-experts to judge the data quality. Transparent and open communication, such as published robustness checks and codes, would counteract these issues. On another note, the second study led to food for thought relating to test and questionnaire design. It suggests that it could be viable to not use a full background questionnaire design in order to free up time and space for additional topics in the questionnaire or cognitive items.

Contents

List of figures

List of tables

1 International large-scale assessments, the Programme for International Student Assessment, and the associated methodology: A brief introduction

1.1 International large-scale assessment and the Programme for International Student Assessment

As early as the 1950s, researchers were contemplating how to measure the quality of an educational system, make comparisons with other countries' systems, and use these results to determine factors that foster student achievement. Even then, generalisations were being made about the inputs and outcomes of different national educational systems. Yet concepts such as internationally valid standards in education were still to be defined. In 1958, a meeting of educational psychologists, psychometricians and sociologists at the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute of Education aimed to address such issues and as a result large-scale assessments were born: The idea of an international study to measure educational outcomes of students and their determinants emerged. Subsequently, a pilot study was commissioned to evaluate the feasibility. This led to the first international large-scale assessment (ILSA) – a mathematics study involving 12 countries – which was conducted in 1961 (Husén & Postlethwaite, 1996).

ILSAs have come a long way since then but the aim has stayed the same – to obtain a valid measure of student achievement in an international context and determine factors within and between educational systems that help or hinder students. In the 1990s and 2000s, a new generation of ILSAs began. The number of participating countries and students increased substantially, but also the aims, structure and organisation of the studies changed (Addey et al., 2017; Howie & Plomp, 2005). Among this new generation of studies is the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) as well as the Trends in International Mathematics and Science Study (TIMSS) and the

Progress in International Reading Literacy Study (PIRLS) of the International Association for the Evaluation of Educational Achievement (IEA). These are some of the most, if not the most, prominent contemporary studies in education. Importantly, the different ILSAs have different emphases in terms of topics and target populations (Fischman et al., 2019; Howie & Plomp, 2005).

PISA was administered in 2000 for the first time and has been repeated every three years since then. It aims to measure the mathematics, science and reading skills of 15-year-olds across countries all around the globe and over time. Over the years, PISA has experienced substantial growth as many (non-OECD) countries have decided to join. Participation numbers rose from roughly a quarter million of students in 43 countries and economies in PISA 2000 to over 600,000 students from 79 countries in 2018 (OECD, 2019b) – making PISA the largest ILSA of its kind (e.g. 49 countries in PIRLS 2016; Mullis et al., 2017, and 64 countries in TIMSS 2019; 2020). Another aspect that sets PISA apart from other ILSAs is its content focussing upon competencies in the domains which (the OECD deems) are relevant for life, rather than specific curricula. Furthermore, PISA's developmental origin is different to the other ILSAs. Multiple governments within the OECD initiated PISA with the explicit aim to inform and serve their policy interests (Schleicher, 2000). In contrast, the IEA, which conducts most other ILSAs, evolved from the meeting of researchers interested in understanding education (Husén & Postlethwaite, 1996).

PISA's aim remains in the present day. The following quote of former OECD Secretary-General Angel Gurría

> *PISA is not only the world's most comprehensive and reliable indicator of students' capabilities, it is also a powerful tool that countries and economies can use to fine-tune their education policies…That is why the OECD produces this triennial report on the state of education around the globe: to share evidence of the best policies and practices, and to offer our timely and targeted*

> *support to help countries provide the best education possible for all of their students. (Schleicher, 2019, p. 2)*

highlights that it is still one of the key goals of PISA to support educational policy-making. The OECD has also had considerable impact in this area. From its origin, PISA has influenced international and national debates and education policies as well as daily school life. One of the first and most exemplary instances happened in Germany after the first cycle. The poor academic performance of Germany in comparison to the other countries came as a surprise to many stakeholders, leading to the so-called 'PISA shock' in Germany. In an attempt to improve the German educational system and PISA scores, extensive reforms to the German school system and curriculum were introduced (Ertl, 2006). Lingard stresses in Addey et al. (2017, p. 447) 'that this kind of reaction is not the exception'. For example, PISA results also led to major educational reforms in Denmark (Egelund, 2008), other European countries (Grek, 2009) and Japan (Takayama, 2008), amongst others.

While discrepancies between national expectations and the PISA results can create opportunities for (sometimes needed) reforms, its influence does not end there. As the results get large amounts of attention in media and politics, countries start to (indirectly) compete for top performances and rankings in the league tables (Addey et al., 2017; Steiner-Khamsi, 2003). As a result, countries that consistently perform well are considered 'reference societies', from which other countries borrow policies and ideas (Steiner-Khamsi & Waldow, 2012). While this can have positive effects on educational systems and lead to improvements, it can also be used to 'scandalise' (bad) results and shape national political landscapes (Addey, 2015; Steiner-Khamsi & Stolpe, 2006). Thus, PISA has become an important tool in and source of soft (educational) governance. Over time, the OECD has thereby changed its role from a mere provider and analyst of data to a political actor who sets the educational discourse and provides guidance (Addey et al., 2017; Bloem, 2015).

As a result, while PISA might be a low-stakes assessment for students, as there is no feedback to schools, teachers or the students themselves, it is high-

stakes for countries (particularly governments and policy officials) due to the widespread public scrutiny of the results. PISA thus, in turn, also has substantial influence on countries and their educational policies. By now, PISA should be understood as a 'social phenomenon and political project' (Sjøberg, 2017, p. 17), which shapes the political landscape and discourse in education. By releasing reports, including or excluding topics as add-ons and focusing on certain results, PISA influences the perception of education, the politics behind it and its value. This is especially true, as it is commonly seen 'as a reliable instrument for benchmarking student performance worldwide' (Breakspear, 2012, p. 4).

Communication from PISA officials and the OECD – both to more academic and boarder audiences – tries to establish that focusing on the national system alone and its improvement is longer enough. Instead countries should focus on adapting policies and characteristics from high-performing countries in order to perform better internationally (e.g. Schleicher, 2000 for expert audiences, 2013 for general audiences - TEDGlobal talk). Even though PISA itself might 'just' be a snapshot of the reality in different educational education systems at a specific point in time, it defines the reality and desired shape of education at the same time (Sjøberg, 2017). But the OECD is not even the only player in shaping PISA and its vision of education. For example, Pearson Inc. – the world's largest commercial education company – was hired and tasked to develop the PISA 2018 framework and therefore what would be measured in that cycle (Molnar, 2014; OECD, 2019a; see foreword).

Yet, it does not end there. The OECD is expanding the 'PISA franchise' and therefore its influence. With PISA for Development (PISA-D), the OECD addresses low- to medium-income countries, which are not included in PISA. They offer help in building testing capacities and conducting PISA-D, but thereby require that it can be linked up with national assessment while aligning PISA-D with PISA. This allows the OECD and PISA to implement and establish assessment cultures and standards in these countries (Addey, 2017). PISA for schools on the other hand is a commercial product where schools can (voluntarily) pay to participate in an additional assessment which

compares them against the PISA main results. This enables a direct link of OECD's global policies into local schools (Lewis, 2020). But the OECD has also recently broadened its concept and impact across age groups with the International Early Learning Study (IELS, also called 'baby PISA') covering 5-year-olds (Auld & Morris, 2019) and Programme for the International Assessment of Adult Competencies (PIAAC, the 'adult version' of PISA; OECD, 2019d) assessing adults between 16 and 65 years.

Overall, the influence and reach of PISA and the OECD is impressively large and prominent. While single students might not be affected, the stakes for countries and its impact are clearly noticeable.

1.2 The basics behind estimating student achievement in ILSAs

As shown above, PISA scores and results can have a large influence on education politics and daily school life around the world. If such far-reaching decisions are made based on PISA scores, it is essential that these scores are carefully computed, sound, and validly comparable across countries and cycles.

If one is unfamiliar with the workings of ILSAs, which stakeholders and applied researchers often are, it is easy to assume that the estimation of student achievement scores is straightforward – such as tests during school, i.e. everyone gets the same test with the same questions at the same time and, in the end, a sum score is formed in order to derive a grade. While this is a valid form of assessment in some situations, due to the nature of ILSAs they require a more complex set-up with extensive statistical and psychometric methods involved (von Davier et al., 2014). The two main drivers for this need are (i) that ILSAs aim to assess students from many diverse countries and (ii) that time for the assessment is limited and attached to large effort, especially for at least three domains at once.

Regarding the first driver, it is impossible to test all students – in the case of PISA, all 15-year-olds in school – in all participating countries. The sheer amount of work and organisational effort would not be feasible. As a result,

schools and students are sampled in each country according to explicit and implicit stratification variables in order to get a much smaller but representative sample of the target population (OECD, 2014b). The aim is to have no substantial differences between the sample and the underlying population, so that cross-country comparisons can be validly made based on the PISA scores. Thereby, sampling criteria need to be comparable in all countries as well as the response rates, which includes schools agreeing to participate, students showing up and actually answering the questions they are presented among others (Micklewright et al., 2012; Rust, 2014). Additionally, working with so many diverse countries entails other issues which need also to be carefully considered. Most importantly, all items and questions should operate the same in all participating countries, i.e. they should capture the same construct or skill cross-culturally in a comparable way. This involves many facets. For instance, everything needs to be translated into the respective languages, no unfair advantages in solving items should be present between countries and variables, especially the background questions, need to be (coded) on an international and comparable scale (Behr & Zabal, 2019).

While the first driver shapes the sampling and shows the need for checks of the comparability of populations, the second driver – time for and effort of conducting the student assessment – impacts the way that the test is designed and how the student achievement scores are computed. For perfect judgement of their knowledge, it would be necessary to ask students all questions relevant to a domain. Yet this is not possible. PISA is a low-stakes assessment that is conducted during school time. Usually, two hours are allocated to testing for PISA (OECD, 2014b). The aim is to assign achievement scores in mathematics, reading and science to all students. It is not possible for all students to be asked many questions across all domains and sub-domains in two hours, so students are randomly presented a fraction of all questions. Information about the parts which they did not answer is estimated based on those that they did answer. Students are administered items based on a systematic design, the so-called rotated test design. In PISA 2012, items within each domain were grouped into item clusters. Four of the 13 item

clusters were then systematically combined to form test booklets. Each student is randomly administered one booklet (OECD, 2014b). The resulting systematic missing data (of the not-administered item clusters) is not problematic per se but requires the usage of different statistical models in order to derive sound and valid scores which are comparable across populations.

The estimation of student achievement scores is a multi-step procedure using all available information about the students, i.e. responses to the cognitive items but also information from the background questionnaires. Due to the rotated test design, students did not answer all questions and not even necessarily questions in all domains. This results in large amounts of missing data by design for each student, sometimes even a complete missing domain by design. Yet there are models which can handle this situation (at least when taken at face value) and estimate achievement scores for all students in all domains. In ILSAs, this is usually done via the so-called 'conditioning model' – a mixture of an Item Response Theory (IRT) model and a latent regression model. In the first step, all cognitive data is used in an IRT model to estimate a multi-dimensional achievement distribution. Subsequently, a latent regression model is conducted to adjust these distributions for population characteristics using students' background information, such as gender, age and socio-economic status. While it may seem counter-intuitive for such background data to be used in the estimation of student achievement, it is unproblematic as long as the scores are not used for the individual but at higher levels, such as group- and country-level averages. Indeed, it is necessary for the computation of valid group measures to counteract attenuation bias (Mislevy, 1991; Mislevy et al., 1992).

The gist behind the imputation of these scores can be explained using the following simplified example. Imagine that two different tests are handed out in school. Half of the students receive only mathematics questions whereas the other half also receives reading questions. Now assume boys and girls perform equally well in mathematics, but there is a gender difference in reading. Female students outperform their male counterparts by 10 points. If the estimation does not account for gender, girls who did not answer the

reading questions would be assigned the same reading scores as boys of the same mathematics performance. When one then estimates gender differences in reading achievement across the whole sample, one would find a difference of just 5 test points rather than 10. As a result, such test designs require that the achievement scores are adjusted for population characteristics in order to produce unbiased, unattenuated group analyses.

While the results of tests in school is usually a single number, i.e. a sum score or percentage, the conditioning model does not directly return a single number or even an estimate in that form. Instead, an achievement distribution is returned, which reflects where the true student ability is most likely located. There are different ways to derive estimates for student achievement from this distribution. For instance, a simple estimate would be the mean of the distribution. In ILSAs, similar to the better known multiple imputation methodology (Rubin, 1987), for each student and domain multiple values, known as plausible values (PVs; five in PISA 2012), are randomly drawn from the achievement distribution (von Davier et al., 2009). In contrast to the mean, PVs are not only estimates for achievement, but also reflect the uncertainty of the estimate. For example, the uncertainty of a student who answered questions in the domain will be lower than the one of a student who did not answer a single question in the domain and where the achievement is solely based on their performance in other domains and background data. Thus, the achievement distribution of the second student will be broader and the PVs will be more spread out. As a result, the dispersion of PVs allows us to draw inferences about the uncertainty of the achievement estimates. If properly accounting for the PVs in the analyses, this is valuable additional information to the achievement estimate (Rubin, 1987).

Overall, it can be said that the process behind the final ILSA scores is a complex procedure with many diverse statistical and psychometric aspects and subtleties. Yet there is little to no awareness of this in policymaking and most areas of research that use these scores (e.g. Jerrim et al. (2017) highlight in the appendix how seldomly methodological aspects are accounted for in economics) – even though all of this shapes the scores and in turn the results and country rankings based on them. As a result, it is of highest importance

that research is conducted to ensure the scores and results based on them are valid and comparable.

1.3 Previous work surrounding psychometric and statistical issues in PISA

Indeed, PISA does not only find favour but also receives criticism, especially from academia. There is an ongoing and active debate about the general idea behind PISA and its usage as well as the methodology behind it (see Zhao, 2020, for synthesis). Psychometric and statistical issues have been investigated in varying forms and degrees of detail by many different researchers. Thereby, some aim to provide overviews or summaries of the general situation or multiple issues, whereas others focus on one specific issue to conduct in-depth investigations. For instance, Hopmann et al. (2007) dedicated a whole book to the question of whether PISA delivers what it promises, where the single chapters deal with different issues. Eivers (2010) gives a short overview of different issues relating to representativeness and cultural fairness in PISA with the help of brief explanations and examples.

Over time, as more research has been published, syntheses of this research were also conducted. After a meta-evaluation of different studies looking at methodological concerns, Fernandez-Cano (2016) concludes that 'PISA is undoubtedly an evaluative undertaking that generated a wealth of research but PISA needs to exercise greater methodological rigor and state its methods clearly in future technical reports' (p. 11). A systematic review reached similar conclusions as it pointed out different technical issues, such as sampling and scaling which threaten the validity of the PISA results – 'structural weaknesses and cracks in the foundations of ongoing PISA foundations' (Hopfenbeck et al., 2018, p. 347).

The regarded issues include but are not limited to sample representativeness, non-response rates and cross-cultural comparability. For example, problems with the sampling and response rates can bias the results and leave cross-country comparisons invalid. Anders et al. (2021) highlight how PISA results and rankings can be related to representativeness. In the case of Canada, low response rates paired with high exclusion rates (most likely) biased the results

in favour of Canada in comparison to the other top-performing nations. Likewise, Freitas et al. (2016) show that the sample in Portugal was not representative of the target population. Similar issues related to sampling can also be found in other countries, such as South Korea, England and Ireland (Eivers, 2010; Micklewright et al., 2012). But even if the response rates are acceptable, this is not necessarily a certain indicator that no bias is involved (Micklewright et al., 2012).

Other issues deal with cross-cultural comparability, such as differences in translation, context and unfair advantages. For instance, test length in reading can vary substantially between countries, e.g. the reading test is 18% longer in German than in the original English version (Eivers, 2010). Furthermore, item difficulties and demands can be perceived differently depending on the language (El Masri et al., 2016). Overall, a substantial amount of studies showed measurement non-invariance or differential functioning of items in cognitive and non-cognitive domains between countries (Hopfenbeck et al., 2018). As just seen, not only the cognitive but also the non-cognitive data suffers limitations and issues. Rutkowski & Rutkowski (2010) show the importance of background variables but also how they are impacted by missing data, misunderstanding of questions and low reliability. This led a call for improvement of background data and better non-cognitive scales (Avvisati et al., 2019; L. Rutkowski & Rutkowski, 2010).

Another aspect that has a high impact on the PISA scores is the conditioning model, a combination of an IRT model and latent regression. It is one of the key foundations of PISA, as it is crucial for the estimation of student achievement scores. Yet only two of 19 chapters in the official technical documentation are dedicated towards it (OECD, 2014b) and it is not explained or even mentioned in the international report at all (OECD, 2014a). This is also reflected in the comparatively sparse literature. Goldstein (2017) states that the complexity of the model and the high level of required expertise lead to high opacity and a threshold for scrutiny and research.

While there are multiple papers on the IRT part of the conditioning model, research on the whole conditioning model is much rarer. In general, key

assumptions of the IRT model are being challenged. Wuttke (2007) questions the unidimensionality of the different domains and whether all items function the same across all countries in PISA. Kreiner & Christensen (2014) go one step further and investigate if results based on the used IRT model, the so-called Rasch model, can be compared sensibly at all. Model misfit and substantial differential item functioning thereby threaten the validity of the PISA results and cross-country comparisons. On the other hand, other research shows that alterations to the IRT model do not largely impact the country comparisons (Jerrim et al., 2018). Meanwhile, Rutkowski (2014) investigated the complete conditioning model with the help of a simplified simulation. She found that the model is sensitive to the misspecifications and systematic error in the background variables can lead to substantially biased country measures.

Yet all those single issues cannot give a complete picture on their own, as they are generally regarded as isolated – even in work summarising or investigating multiple aspects. They are intertwined and impact each other. For example, if an item has a non-optimal translation, it can be biased and impractical for later secondary analyses, but before it is also used in the conditioning model and can thus introduce bias into the achievement scores (L. Rutkowski, 2014). Therefore, it is vital to have a realistic overview of all potential issues with the data. The Total Survey Error (TSE) framework considers the whole process behind surveys in general and aims to investigate all potential sources of error. Thereby, it encompasses 'the entire set of survey design components that identify the population, describe the sample, access responding units among the sample, operationalise constructs that are the target of the measurement, obtain responses to the measurements, and summarise the data for estimating some stated population parameter' (Groves & Lyberg, 2010, p. 850). The underlying aim is to gauge the quality and meaningfulness of the survey and identify improvements, if possible. Schnepf (2018) discusses why this is not only important for surveys in general but also for ILSAs and PISA. Thereby, examples for the different error components are showcased in order to raise awareness for the fact that the TSE is unlikely to be negligible, which in turn raises doubts about the current usage and

interpretation of the PISA data and results. To my best knowledge, I am not aware of any published work apart from Schnepf (2018) that investigates the TSE in PISA or even attempts a broader and more comprehensive investigation of potential error sources in PISA.

1.4 Thesis outline

This thesis is structured in five main parts. In this first chapter, the introduction, I outline the motivation and relevance of my research.

Chapter 2 focuses on one part of PISA's student achievement computation, the conditioning model, where background variables are used in combination with the cognitive items to derive student achievement scores. Thereby, the aim is to investigate the impact background variables can have upon the PISA results. As a starting point, the PISA 2012 PVs were replicated as closely as possible to allow conclusions and references to PISA in further analyses. The resulting model was then systematically varied by including and excluding different sets of background variables in its specifications. Differences in scores were analysed to gauge the impact that this had upon the relative position of countries in the PISA rankings. The key finding is that the exact specification of the conditioning model matters; including or excluding variables can substantially change student achievement scores. As a result, cross-country comparisons of PISA scores, such as country rankings, can change quite dramatically depending upon the statistical methodology used. This is most pronounced for achievement scores in the minor domains (reading in my empirical application) and for cross-national comparisons of educational inequality.

Chapter 3 builds upon the foundations laid in the second chapter: A simulation study was conducted to further investigate some of the research questions relating to the impact of background variables in student achievement estimation. But additionally, the simulation is used to examine the impact that other key aspects surrounding the plausible value computation have on country averages and inequality measures. The simulation is based upon the design of the PISA 2012 assessment and allows me to investigate some research questions which would not be possible with real-life PISA data

only. Thereby, I systematically vary the variables included in the conditioning model, how these are prepared, the selection of test questions that students were required to answer and whether the background questionnaire has a full design. This study confirms that specification of the conditioning model matters. The presence of a single variable can introduce bias into the results, especially in the minor domains, though this impact can be reduced by asking students test questions in all subjects (something not done in PISA). In contrast, how the conditioning variables are prepared and whether a full background questionnaire is used or not has little to no impact upon the results.

While the previous two chapters already suggest that country measures can be impacted through multiple pathways, it is the explicit aim of Chapter 4 to provide a broader and more comprehensive review. It aims to show different psychometric and statistical properties of PISA and their corresponding issues, by using a case study, exploring the link between highest parental education and achievement scores in Germany in PISA 2012, in order to allow for an in-depth investigation. Overall, I investigated six different psychometric and statistical properties that potentially impact parental education group comparisons in Germany. Thereby, the study shows how diverse and different sources of bias can be in PISA, how they are intertwined and that they indeed affect the validity of the results. While not all investigated issues introduce bias into the scores and results, multiple aspects did.

This thesis closes with conclusions in Chapter 5. Thereby, the key findings from Chapters 2–4 are summarised first. Subsequently, the different findings are set into reference of each other and the general theme in order to highlight the overarching contribution of this thesis. Furthermore, limitations that apply to either single chapters or the complete thesis are discussed critically. While this thesis led to new knowledge and added to the literature, it has also raised further research questions specifically connected to the psychometric and statistical properties in ILSAs. So, to bring this thesis to an end, potential opportunities for future research based on this thesis are shown.

1.5 Contribution of this thesis

Subchapter 1.3 illuminates the situation of research regarding ILSAs and related methodological issues. Diverse aspects can threaten the validity and comparability of PISA scores and results. Thereby, it is natural that some areas receive more attenuation than others. With the help of Chapters 2, 3 and 4, I intend to address gaps in the literature so far and make valuable contributions in this area of research.

Chapter 2 aims to shed further light on the conditioning model. While there is literature about the conditioning model, the theory behind it and its benefits (e.g. Khorramdel et al., 2020; Mislevy, 1991; Mislevy et al., 1992), little work has been published scrutinising the usage of the conditioning model in ILSAs. Due to the complex nature of ILSAs, the model and the large sets of items and questions, research investigating the conditioning model in ILSAs requires great amounts of computational effort and time. For instance, Rutkowski (2014) looked at conditioning in ILSAs, but through a simplified and reduced simulation in order to highlight how bias in the conditioning variables could potentially bias ILSA results. In comparison, this chapter substantially adds to the literature, as

- it scrutinises the conditioning model in a realistic PISA setting. This means that the actual PISA 2012 data for all countries is used, the methods are replicated as closely as possible and the same items and questions are included in the model as in PISA.
- my key finding shows that key measures and country rankings can change substantially depending on the exact model specification.
- the OECD does not publish its code for the PV estimation, but I make my code publicly available (Zieger, 2021) for researchers to scrutinise and reuse.

As a result, this chapter does not only fill a gap in the literature but is relevant in general, as little rationale and explanation behind the conditioning model specification and its impact are communicated in the official PISA reports.

Chapter 3 also addresses the lack of critical literature about the conditioning model in ILSAs. Thereby, it starts by further investigating some aspects from Chapter 2 but also assesses the impact that properties (indirectly) affecting the conditioning model can have on country measures. This is done via a simulation study that replicates the PISA data and design and enables the investigation of research questions that would not be possible otherwise. Consequently, the primary contributions to the literature of this chapter are:

- To provide the first evidence on the impact of conditioning variables and design characteristics on plausible values in a real-life setting. This is important as no official (in-depth) robustness checks and analyses are publicly available for PISA.
- It provides further evidence (over and above that presented in Chapter 2) that the conditioning model specification matters. The inclusion or exclusion of one single variable can bias the results.
- I find that this bias can be reduced significantly if all students were administered questions in all domains, which PISA does not do. This is an interesting finding, as most other ILSAs (e.g. TIMSS) administer questions in all tested domains to all students.
- I show – for the first time – that some aspects of the conditioning model have little to no impact upon the PISA results, such as the background questionnaire design, i.e. if all students answered all questions. This finding is new and relevant, as it opens test design options, especially in combination with the previous finding.

Thus, this chapter contributes both to the academic and general debate about ILSAs and how students should be tested as well as how their achievement scores are estimated.

Most work surrounding methodological issues in ILSAs focuses on a single isolated aspect, like Chapter 2, or also includes a few related properties, such as in Chapter 3. Yet those issues do not occur separately in reality and regarding them (semi-) isolated does not give the full picture. ILSAs have numerous steps and properties which all affect and relate to each other. The TSE framework acknowledges this and provides a guideline on which aspects

of surveys can potentially be problematic and should therefore be assessed. Schnepf (2018) highlights the need for transparent evaluation and communication of the aspects of the TSE framework in ILSAs, but little to no empirical research has been done regarding PISA and the TSE framework. Chapter 4 therefore contributes to the existing literature by:

- Comprehensively investigating diverse ways how country measures can be biased outlined in the TSE framework.
- Identifying six statistical and psychometric aspects, including the influence of background variables via the conditioning model, that can potentially impact group comparisons.
- Illustrating how diverse and intertwined sources of bias are in PISA, which is also a valuable lesson for users of PISA data and results that can be applied to other situations.

Overall, this thesis sheds light on the details of how student achievement scores are computed, how different psychometric and statistical properties impact student scores and in turn the validity and comparability of key measures. Thereby, the overall contribution to the academic literature is twofold:

1. The conditioning model is scrutinised closely and the indirect impact of background variables through the achievement score computation is highlighted.
2. The relationship between background variables, achievement scores and statistical and psychometric properties of ILSAs is investigated on a more comprehensive level in order to identify potential issues and bias in country measures.

All three research projects in this thesis identify points of concern in the process. Yet the main contribution to the literature does not lay in the sole identification but its consideration from different points of view and in relation to the total picture. Furthermore, this thesis does not only add to the scientific debate but also two further general contributions emerge:

3. This thesis aims to highlight the dependency of country measures on different statistical and psychometric properties. Limitations and implications should be better communicated to researchers, stakeholders, and the media.
4. The need for better transparency and official documentation is addressed. It becomes clear that a lack of details and adequate (and consistent) documentation remains a problem with PISA.

2 Conditioning: How background variables can influence PISA scores

2.1 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 concern. As discussed by Rutkowski and Rutkowski (2016) and others (Gillis et al., 2016; S. Hopmann et al., 2007) this includes issues such as sample representativeness, non-response rates, population coverage and cross-cultural comparability. For instance, in the case of Portugal, Freitas et al. (2016) found substantial differences between the target population and the sample which may have introduced bias into the results. Other countries, such as South Korea, England and Ireland, have also experienced questionable movements in PISA scores over time, potentially due to sampling issues (Eivers, 2010; Micklewright et al., 2012). Other criticisms of PISA include potential bias

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

However, perhaps the most controversial element of PISA is the scaling model 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, 2014b). 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 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' (PISA 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 et al., 1992). Second, by conditioning upon pupils' background characteristics, the precision of population estimates should be enhanced (e.g. smaller standard errors in average PISA scores; van Rijn, 2018). On the downside, this adds substantial complexity to the generation of PISA scores, leading to the criticisms that it is opaque.

While conditioning upon background characteristics is a key part of the production of PISA scores, only two out of nineteen chapters of the PISA 2012 technical report are dedicated to the computation of plausible values (OECD, 2014b). This highlights the lack of examination of the topic, which is also evidenced by the scarcity of research conducted on this matter in international large-scale assessments (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 motivate its use (smaller standard errors and more accurate estimates of covariation with background characteristics)? Or does it simply add a great deal of complexity for little discernible gain?

This chapter aims to answer such questions about the conditioning model used in PISA and fill the gap in the literature. It begins by investigating how closely the PISA plausible values can be reproduced using publicly available documentation about the procedures used. I then compute alternative plausible values using different variants of the conditioning model. Results from using the full conditioning model are then compared to those using only basic parts of the model, and 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 allows close replication of the plausible values in the major domain (mathematics in the PISA 2012 data I 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 an impact in some of the minor domains (particularly reading). In other words, different versions of the conditioning model can lead to different country-level PISA scores. Third, there is evidence that the specification of the conditioning model can have substantial, but not necessarily predictable, impacts upon important measures of educational inequality. Finally, I 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).

This then leads me 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, I conclude that the OECD should publish more sensitivity analyses around the conditioning model and make more detailed information about their methodology publicly available.

2.2 Methods

2.2.1 Data

In this chapter, I 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, 2014b, 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 subpopulation (within-country) comparisons (OECD, 2014a, p. 256). The average school and student response rates after replacement are 98% and 92%, though there are substantial differences between countries. Overall, PISA 2012 encompasses 478,413 students in 64 countries and economies (Cyprus excluded[1]).

2.2.2 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

[1] Not all participating countries recognised Cyprus as an independent country. Data for Cyprus was not published.

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.[2] 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, 2014b, pp. 30, 31). Therefore, complex techniques (IRT and latent regression) are used to impute data in domains where students have not answered any test questions (e.g. reading) from how they performed upon test questions in other domains (e.g. mathematics and science) and their background characteristics (e.g. gender, socio-economic status, attitudes towards mathematics, enjoyment of school). See OECD (2014b, 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 is only available for a subset (OECD, 2014b, p. 58). In addition to the mandatory questionnaires and domains (student and school questionnaires and the mathematics, reading and science test), countries could also administer some optional elements of PISA. This included parental, educational career and information communication technology questionnaires as well as additional assessments in digital reading, computer-based mathematics, financial literacy and problem solving (OECD, 2014b, pp. 22, 259, 260; see also Appendix A.1 for details). The additional domains were computer-based assessment, while the core domains were paper-based.

[2] Each cluster contained 30 minutes of test material. Two of the mathematics 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.

2.2.3 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 plausible values (see Chapter 9 and 12, especially pp. 159, 253, 254 of OECD, 2014b).

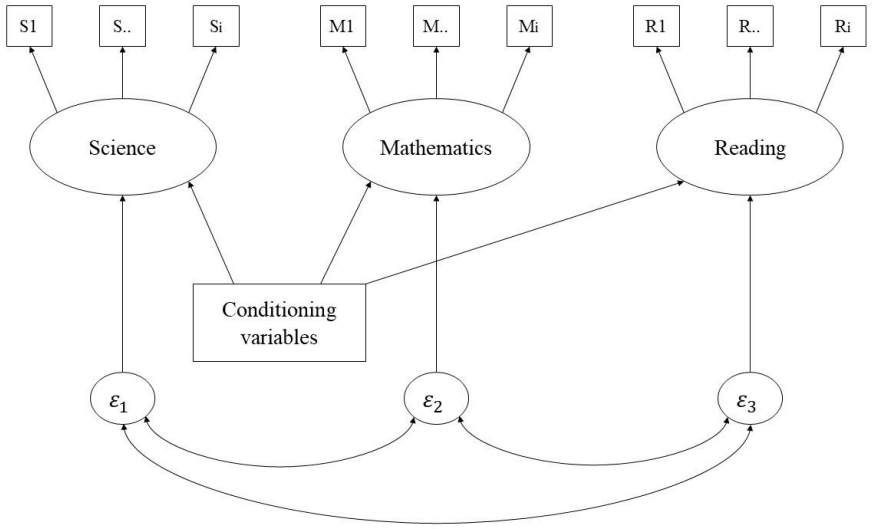
- First, for each core domain (reading, mathematics, and science) the item difficulties are determined using a common sample[3] via IRT model. These are then fixed for all later stages.
- Second, responses to the background questionnaires are recoded for each country. These are then used as 'conditioning 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 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 students' achievement in each subject. A simplified illustration of the model used can be found in Figure 2.1. 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 in their estimated reading, science, and mathematics achievement.
- 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).

[3] The common sample consists of 500 students from each country, except for Liechtenstein, where students were randomly selected (OECD, 2014b, p. 233).

- Finally, these plausible values are transformed by common item equating onto the PISA scale. This final element facilitates comparisons of PISA scores over time.

The focus of this chapter is the role of the conditioning model (i.e. the use of school and student background data) detailed in the third bullet point above[4].

Figure 2.1 Simplified illustration of the PISA scaling model used to generate the plausible values



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.

2.2.4 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, 2014b, p. 145). In a nutshell, they are essentially an application of Rubin's (1987) well-known multiple imputation (MI) methodology applied

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

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)[5] – see (Mislevy, 1991; Mislevy et al., 1992) and (b) reduce uncertainty in measurement (van Rijn, 2018).

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 5 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

[5] 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.

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).

2.2.5 Replication of the PISA methodology

In order to investigate how the specification of the conditioning model influences PISA results, I 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, 2014b, pp. 144–146), let:

- the items be $i = 1, \dots, I$ and
- response category $k = 0, \dots, K_i$ with $K_i = 1$, if item i is binary or $K_i = 2$ if item i has partial credit.
- The value vector is denoted as $\boldsymbol{X_i} = \left(\boldsymbol{X_{i1}}, \dots, \boldsymbol{X_{iK_i}}\right)^{\boldsymbol{T}}$ with $X_{ij} = 1$, if the value of item i is in the respective category, otherwise 0.
- Let $\boldsymbol{\xi}^T = \left(\boldsymbol{\xi_1}, \dots, \boldsymbol{\xi_p}\right)$ be the p item parameters,
- $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)$ denote the latent variable of the D domains,
- $\boldsymbol{a}_{ij}, (\boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{I}, \boldsymbol{j} = \boldsymbol{1}, \dots, \boldsymbol{K_i})$ of length p denote the design vectors in the IRT model which form design matrix $\boldsymbol{A}^T = \left(\boldsymbol{a}_{11}, \dots, \boldsymbol{a}_{1K_1}, \boldsymbol{a}_{21}, \dots, \boldsymbol{a}_{2K_2}, \dots, \boldsymbol{a}_{IK_I}\right)$,
- $\boldsymbol{b_{ijd}}$ denote how response category j of item i loads onto dimension d, which is combined into $\boldsymbol{B}$ as follows $\boldsymbol{b}_{ik} = (b_{ik1}, \dots, b_{ikD})^T$, $\boldsymbol{B}_i = \left(\boldsymbol{b}_{i1}, \dots, \boldsymbol{b}_{iK_i}\right)^T$ and $\boldsymbol{B} = (\boldsymbol{B}_1^T, \boldsymbol{B}_2^T, \dots, \boldsymbol{B}_I^T)^T$,
- $f_\theta(\boldsymbol{\theta}; \boldsymbol{\alpha})$ 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 first state the IRT model used in PISA 2012 and then focus on the conditioning model. The IRT model is based on the concept of a 1-pl IRT model, meaning that only item difficulty is estimated (the discrimination and 'guessing' parameter are held fixed as 1/0). In contrast, to 2-pl and 3-pl IRT models also estimate item discrimination and guessing parameters. Yet, the model is not used in its basic form but in a generalised form, the Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM Adams et al., 1997), to facilitate some characteristics of PISA – multidimensionality and partial credit.

While most items in PISA 2012 have only one correct answer and are therefore binary, a few items (15 out of 206 items in the core domains) were partial credit, i.e. students could gain up to two points in constructed responses items, but also only one if the answered was not exactly as desired, which results in a scoring of 0, 1 or 2. In the partial credit model, the probability for achieving both one and two points is estimated. Furthermore, PISA aims to estimate student achievement in (at least) three domains at once, resulting in (at least) three dimensions.

The MRCMLM can estimate numerous models of different complexity, e.g. it can estimate a simple 1-pl IRT model without any extension. But it is also able to accommodate a 1-pl IRT model which is extended for partial credit and multidimensionality. The probability of scoring j in item i is thereby defined as follows:

$$P\left(X_{ij} = 1|\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}|\boldsymbol{\theta}\right) = \frac{\exp\left(\boldsymbol{b}_{ij}\boldsymbol{\theta} + \boldsymbol{a}_{ij}^{T}\boldsymbol{\xi}\right)}{\sum_{k=1}^{K}\exp(\boldsymbol{b}_{ik}\boldsymbol{\theta} + \boldsymbol{a}_{ik}^{T}\boldsymbol{\xi})}.$$

Where $\boldsymbol{B}$ manages the relationship between items and the dimensions/domains; in $\boldsymbol{b}_{ij}$ is the dimension determined to which the item belongs. $\boldsymbol{A}$ handles the relationship between the item and its model

parameters, in case of partial response $\boldsymbol{a}_{ij}^{T}$ builds the linear combination of respective item difficulties.

Based on this IRT model, the latent achievement of student can be estimated. 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[6]:

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

In the above, no conditioning model has been applied. Now, let's assume that students from different subpopulations (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 is fixed, the mean μ is replaced with the regression model estimate $\boldsymbol{Y_n'\beta}$. As a result, the latent variable is now represented through $\theta_{in} = \boldsymbol{Y_n'\beta} + \varepsilon_n$, 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 I plug this regression into the density function, I 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:

[6] 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.

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

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, I apply this approach to the PISA 2012 data as described in Appendix A.2.

2.2.6 How is 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[7] and then used as conditioning variables ($\boldsymbol{Y_n}$). Within the conditioning model, each background variable is treated as either (OECD, 2014b, p. 157):

- A direct regressor. These are added straight to $\boldsymbol{Y_n}$ without any further processing, just recoding. Only the following handful of variables are direct regressors: gender, school ID, grade, mother's and father's socio-economic index and booklet IDs[8]. These variables are therefore available for all students in the PISA conditioning model[9].
- An indirect regressor. The remaining (vast majority) of background variables are recoded in one of three ways: (a) combined into

[7] By recoding, I 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 for all different categories, I 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 the reference category was selected or 0 if neither). By pre-processing, I 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, 2014b, pp. 157, 421–431).

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

[9] This is true even with the questionnaire rotation used in PISA 2012, as questions capturing this information were situated in the core part, i.e. items were seen by all students.

preliminary questionnaire indices; (b) dummy-coded if categorical or (c) centred and a dummy variable added for missing information if numerical[10]. A principal component analysis is then conducted on these recoded variables, with as many components retained as necessary to explain 95% of the variance. The retained components are then included in the vector of conditioning variables $\boldsymbol{Y_n}$. According to the official documentation, no imputation, or other approaches to dealing with the large amounts of missing background data (due to the rotated questionnaire design) were applied. The conditioning variables $\boldsymbol{Y_n}$ are computed separately by country and may therefore vary (e.g. in terms of the number of components that were retained). For each country, all available information was used[11].

2.2.7 Variations of the conditioning model

After trying to reproduce the published values, I 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, I 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

[10] The exact details for all recoding can be found in Annex B in the technical report (OECD, 2014b, pp. 421–431).
[11] For example, Germany administered the parental questionnaire. This meant that more items were included in the Principal Components Analysis (PCA) for the computation of indirect regressors in Germany than in most other countries.

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 chapter are done within R (R Core Team, 2019) using the 'TAM' (Robitzsch et al., 2018) and 'intsvy' (Caro & Biecek, 2017) packages. Further details about the computational procedures (both the replication and altering the conditioning variables) can be found in Appendix A.3. For the comparisons and analyses of the produced plausible values, I accounted for the sample design by using Balanced Repeated Replication (BRR) weights in combination with the final student weight.

2.3 Results

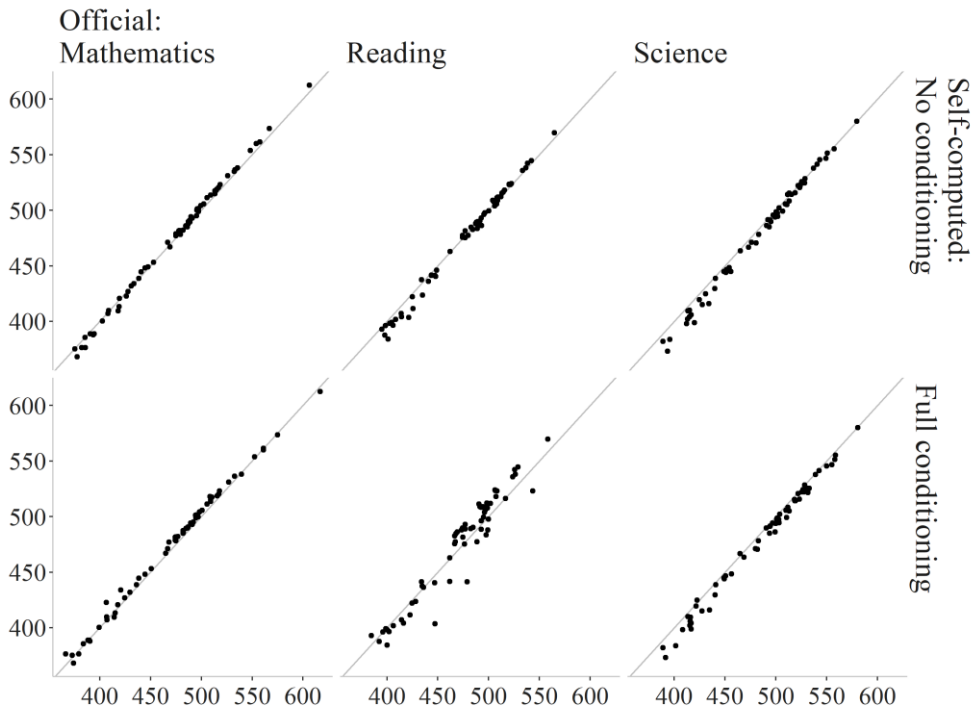
2.3.1 Average scores

Figure 2.2 illustrates the relationship (at the country level) between my self-computed country average PISA scores and the 'official' OECD scores. The upper panel refers to my 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.

My replication of the PISA plausible values has succeeded to different degrees. The correlation between the self-computed country averages and the 'official' country averages is very good for the major domain (mathematics) where correlations are above 0.998. Similar results hold for science (one of the minor domains). Although there is slightly more variation between the official country average science scores and the replicated values, the cross-country correlation in the results is still strong; the Pearson correlation is .996

with full conditioning and .998 without. In other words, in these two domains, the impact of conditioning upon the results is trivial.

Figure 2.2 Countries' average PISA scores. Official versus self-computed scores



Note: The 'official' country average scores are plotted along the horizonal axis and self-computed scores along the vertical axis. The upper panel refers 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 Pearson correlations, starting in the top, left-hand graph and working right, are .999, .997, .998, .998, .965 and .996.

The results for reading (the other minor domain) are, however, more of a concern. In the upper panel, when no conditioning is applied, the self-computed country averages closely replicate the official OECD scores (Pearson correlation = .997). This changes in the bottom panel once I condition upon background data, with the correlation falling slightly to .965, leading to many countries experiencing an important change to their results. For instance, at the extreme, the average reading score in Chile increases from 441 to 479 (i.e. by more than 0.3 standard deviations), while it falls in the

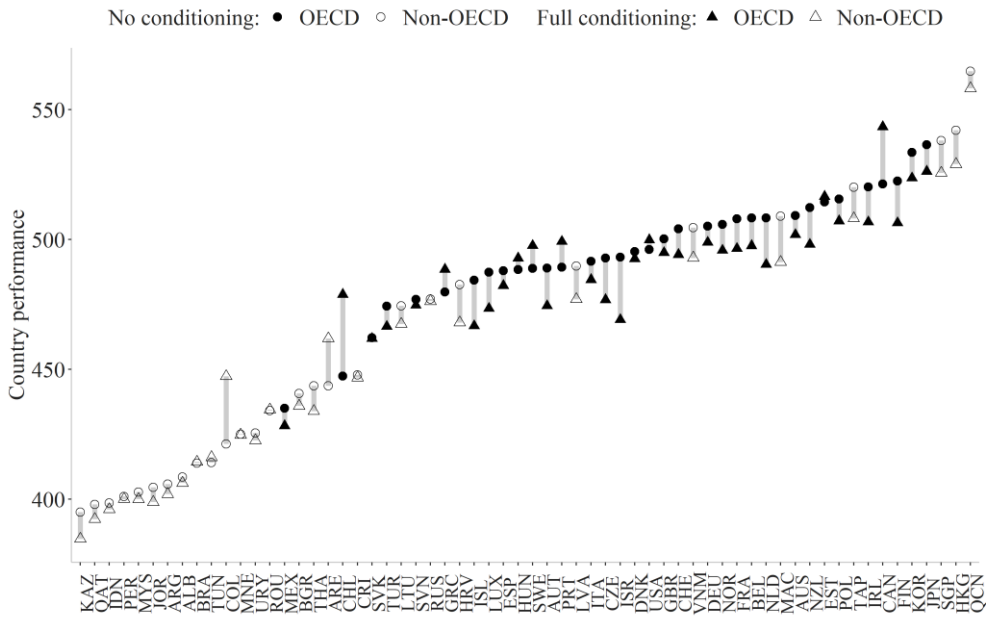
Netherlands from 511 to 490 (i.e. a drop of around 0.2 of an international standard deviation). Indeed, when conditioning upon background characteristics, the self-computed estimates of average reading scores in lower performing countries have a slight tendency to be higher than the official results, while the self-computed average reading scores for high performing countries tend to be slightly lower.

Given these results, from this point forward, I focus mainly upon findings for reading in the main text. Full (or additional) results for all three domains separately can be found in Appendix A.4 (mathematics), A.5 (science) and A.6 (reading).

2.3.2 The impact of conditioning

To illustrate the possible impact of conditioning on average reading scores, I focus on the comparison of the self-computed plausible values with and without conditioning. This can be found in Figure 2.3. The length of the lines depicts the effect that conditioning has on country average reading scores.

Figure 2.3 Country average reading scores with and without conditioning



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

In general, average reading scores within most countries decline when conditioning is applied (triangular markers in Figure 2.3 tend to be lower than the circular markers), with only 13 out of 62 countries experiencing an increase. Indeed, as Figure 2.3 demonstrates, the impact of conditioning in low-performing countries is relatively small (the circular and triangular markers tend to sit on top of each other) while in middle-to-high performing countries the impact of conditioning seems larger (the circular and triangular markers are further apart). Yet, there are some expectations in lower-performing countries like Chile and Colombia, which also experience a substantial impact on their average scores. In terms of the often-cited PISA 'country-rankings', conditioning has relatively little impact upon the composition of the top and bottom performing groups (though with some exceptions). It does, however, lead to important changes around the middle, where country averages are close to each other and changes due to model specification occur in different magnitudes and directions. For instance, Israel drops 13 places (from 25th to 38th) while Portugal rises 15 places (form 29th to 14th).

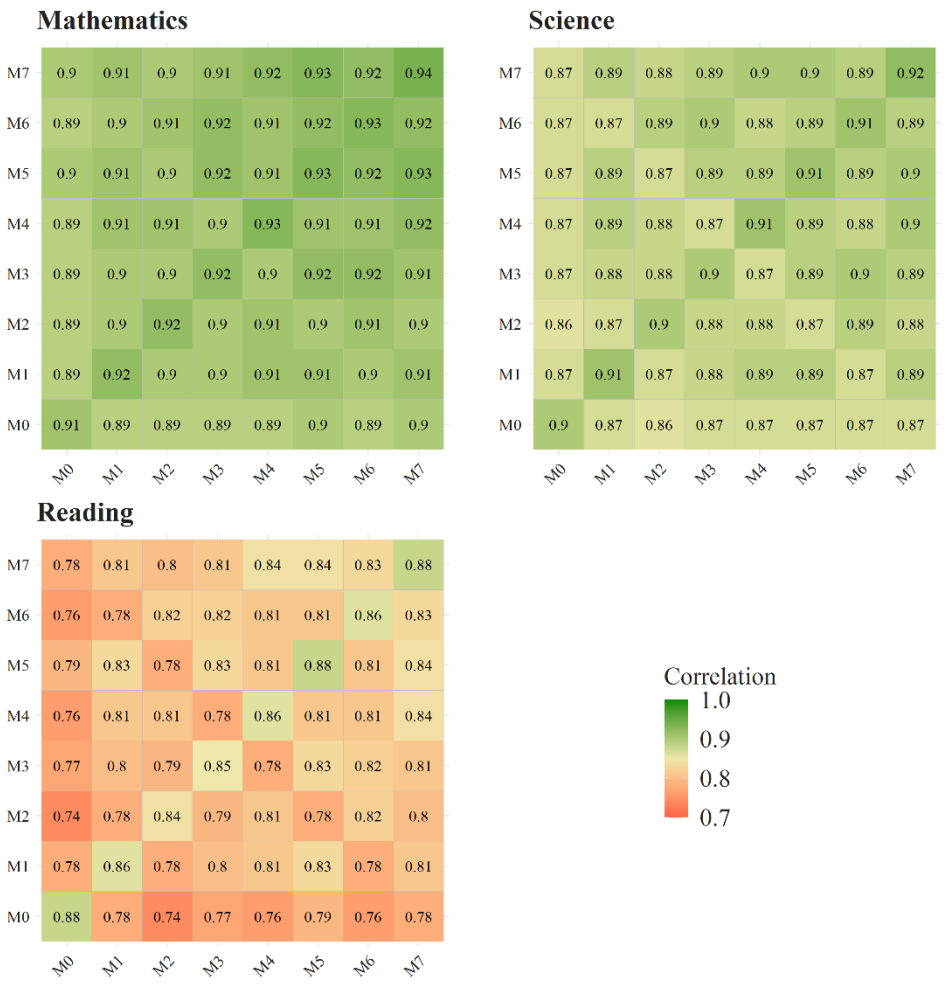
What part of the conditioning model leads to this difference? The next part of the analysis compares results using different specifications of the conditioning model, focusing upon three different subsets of conditioning variables: (a) school direct regressors (i.e. contrast codes for each school); (b)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

Figure 2.4 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 1.0. On the other hand, red shading denotes a correlation of 0.7 (just below the minimum I 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 orange 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 that 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.

Figure 2.4 Correlations of the individual-level plausible values in mathematics, reading and science with different specifications of the conditioning model



Note: 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Second, these findings are reinforced when looking at the diagonals in Figure 2.4. The correlations sit between 0.91 and 0.94 in mathematics, 0.90 and 0.92 in science and 0.84 and 0.88 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.

Overall, in both aspects reading stands out, as its correlations are noticeably lower (more orange) than those of mathematics and science. While it is impossible to say with certainty where this stems from, there are different factors which most likely contribute to it. First of all, reading is a minor domain which means that less questions (and thus less information) is available to compute achievement scores in comparison to mathematics. Yet science is also a minor domain but the correlations are higher – suggesting other factors are also likely at play One possibility is that achievement scores in reading are at least partly derived based on the mathematics and/or science performance of the students. If the items in mathematics are less predictive of reading achievement (than science achievement), then this might also lead to higher uncertainty in the reading scores. Another possibility may be that the reading items are constructed in a particular way, e.g. many questions focussing on one text, which may also lead to more uncertainty and lower discrimination of students' reading achievement.

Table 2.1 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 2.1 Variation in estimated average PISA reading scores by conditioning model specification in the OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Japan	537	526	519	524	530	527	519	526
South Korea	534	542	521	-	523	542	519	524
Finland	523	524	500	524	504	523	504	506
Canada	521	545	535	530	541	537	539	543
Ireland	520	518	516	523	516	523	536	507
Poland	516	529	509	516	508	517	508	507
Estonia	514	529	522	536	519	535	519	517
New Zealand	512	512	498	512	498	511	498	498
Australia	509	501	493	497	500	505	495	502
Netherlands	508	509	488	509	489	509	490	490
France	508	512	491	500	486	510	490	497
Belgium	508	502	499	506	495	510	498	498
Norway	506	492	457	462	467	494	464	496
Germany	505	507	499	509	499	513	496	499
Switzerland	504	508	492	506	494	507	493	494
United Kingdom	500	499	492	500	495	499	491	495
USA	496	510	500	503	503	509	498	500
Denmark	495	489	485	506	488	501	486	493
Israel	493	505	489	498	469	502	481	469
Czech Republic	493	490	478	490	476	489	478	477
Italy	492	491	481	490	483	490	483	485
Sweden	489	519	489	506	492	503	487	498
Portugal	489	526	502	501	502	500	501	499
Austria	489	490	479	489	472	493	473	475
Hungary	488	484	487	489	483	487	491	493
Spain	488	490	479	488	482	489	481	482
Luxembourg	487	488	473	487	473	487	473	473
Iceland	484	484	471	483	466	483	468	467
Greece	480	479	485	479	485	479	489	489
Slovenia	477	484	475	476	472	492	474	475
Turkey	474	474	460	473	465	473	466	467
Slovak Republic	462	479	461	472	472	479	457	462
Chile	447	489	489	494	484	477	479	479
Mexico	435	433	427	432	427	432	429	428
OECD mean	**497**	**502**	**489**	**497**	**490**	**501**	**490**	**491**
OECD median	**496**	**502**	**489**	**500**	**489**	**502**	**490**	**495**
Cor with M0	**1.00**	**0.83**	**0.78**	**0.80**	**0.80**	**0.92**	**0.81**	**0.85**
Cor with M7	**0.85**	**0.90**	**0.91**	**0.85**	**0.96**	**0.91**	**0.92**	**1.00**

Note: Figures illustrate how average PISA reading scores vary depending upon the specification of the conditioning models. Results for non-OECD countries reported in Table A.10. 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 regressors, 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.

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

The average reading scores (and ranking) for selected countries are particularly sensitive to conditioning model specification. For example, the performance for Israel drops substantially when all direct regressors are used as in M4 and M7 (orange cell, corresponding to 30th place). But it displays visibly lighter orange/yellow colour for the other models (between 16th and 19th place for other models with exception of 23rd place for individual direct and indirect regressors). Especially Norway is also salient as the colour change seems highest here. Throughout this chapter I find that Norway is one of the countries with the highest impact of conditioning in reading. Thereby, average scores are consistent in mathematics and science (see Appendix A.4.1 and A.5.1), which highlights that most likely there are much higher levels of uncertainty in reading leading to higher influence of the conditioning model. Reasons for this in comparison to other countries could potentially be that reading does not work as in this language or culture (which is similar to Sweden which also showed noticeable differences). Furthermore, both the

school component and the background components have strong influences but not necessarily in the same way. 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.

2.3.3 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 (e.g. OECD, 2013), and UNESCO uses PISA data for their report on educational inequality (Gromada et al., 2018) as well as in research such as Oppedisano and Turati (2015) and Gamboa and Waltenberg (2012). I therefore illustrate in Table 2.2 how sensitive a widely used measure of educational inequality (the difference between the 90^{th} and 10^{th} 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 2.2 is that conditioning leads to an increase in estimated educational inequality (on average) across OECD countries. Specifically, the average percentile gap rises by 23 points, from 211 with no conditioning to 234 when full conditioning is applied. The gap between the 90^{th} and 10^{th} percentile increases substantially as soon as any conditioning is used.

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

Table 2.2 Estimates of inequality (90^{th}–10^{th} percentile difference) in PISA reading scores across countries by specification of the conditioning model in the OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	161	191	195	192	196	193	197	197
Chile	174	128	127	125	139	149	148	150
Estonia	179	172	192	138	200	164	201	202
Turkey	190	216	234	216	224	218	223	221
Denmark	191	190	176	131	196	193	204	209
Ireland	193	175	193	186	189	189	213	201
Poland	196	240	212	209	212	213	212	212
Spain	199	226	231	227	233	227	233	234
Czech Republic	200	218	226	220	222	217	224	221
Switzerland	204	235	240	231	243	234	239	241
Canada	205	215	243	189	231	200	229	226
USA	207	210	192	174	187	168	183	183
Hungary	208	191	205	203	191	200	204	214
Austria	209	208	234	236	235	231	242	240
Germany	210	198	216	203	216	216	225	230
Finland	213	239	259	235	254	236	251	249
United Kingdom	213	240	240	236	240	241	239	239
Slovenia	213	231	264	242	258	204	265	259
Italy	214	245	252	245	253	245	252	251
Portugal	214	213	206	188	224	205	200	219
Netherlands	215	237	245	240	241	239	245	242
Iceland	216	242	253	241	250	243	249	249
Greece	218	250	257	252	255	251	262	259
Norway	220	204	145	165	187	231	192	244
Australia	223	240	239	240	251	256	249	258
Japan	223	270	277	273	253	263	268	231
Belgium	229	209	187	223	216	226	226	232
Sweden	231	230	244	234	245	235	258	243
New Zealand	238	266	270	265	271	266	270	272
Israel	239	216	300	247	265	227	298	271
France	240	247	260	258	253	255	261	274
Luxembourg	240	268	276	269	277	269	277	277
Slovakia	240	262	258	256	263	261	254	272
OECD mean	**211**	**222**	**229**	**218**	**229**	**223**	**233**	**234**
OECD median	**213**	**226**	**239**	**231**	**235**	**227**	**239**	**239**
Cor with M0	**1.00**	**0.71**	**0.63**	**0.71**	**0.73**	**0.77**	**0.74**	**0.83**
Cor with M7	**0.83**	**0.80**	**0.79**	**0.82**	**0.91**	**0.88**	**0.89**	**1.00**

Note: 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. Results for non-OECD countries reported in Table A.11. 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 regressors, 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 0.79 and 0.91. At the same time, none of the specifications shows a particularly high correlation (r between 0.63 and 0.83) with M0 (no conditioning applied). In general, high variation between the different specifications can be seen through the varying colour patterns.

Finally, no clear country patterns can be identified, either in relation to changes in average scores nor concerning changes between model specifications. Again, Norway has high fluctuation in the level of inequality measure depending on the chosen specification (between 2nd and 24th place). This matches with before, especially as the other two domains mainly experience difference between no conditioning and conditioning which is in line with the general picture there (see Appendix A.4.2 and A.5.2). Whereas other countries, such as Turkey, see a rather constant shift also in reading as soon as conditioning is applied (4th place without conditioning and between 10th and 15th place as soon as conditioning is applied).

When examining the corresponding tables in mathematics and science (Tables A.2 and A.4 for mathematics and Tables A.6 and A.8 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.

2.3.4 The association between PISA scores and background characteristics

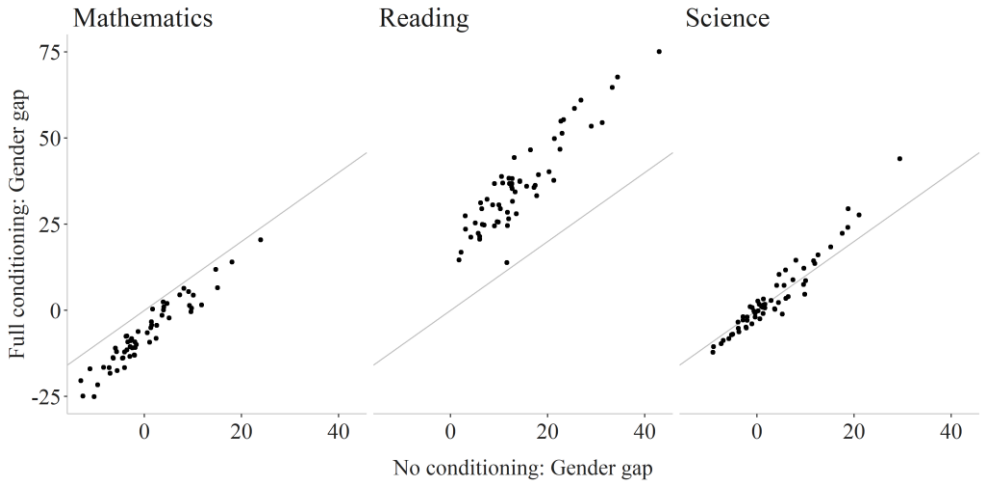
PISA is also often used (including by the OECD) to compare the performance of groups (e.g. gender, socio-economic status, language). But it is well-known that IRT, when used in conjunction with rotated test designs, can lead to attenuation of such group differences (Mislevy, 1991). One of the main motivations for using conditioning models is to counteract such attenuation bias. I 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. three 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.

Figure 2.5 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 2.5 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 36 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 7

points higher than girls.[12] The gender gap is more concentrated in science, with little obvious change occurring when conditioning is used, especially around zero where there is no gender gap to be attenuated to begin with.

Figure 2.5 Country gender gap in mathematics, reading and science with and without conditioning



Note: 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 Pearson correlations, starting left and working right, are $r = .954$, $r = .917$ and $r = .966$.

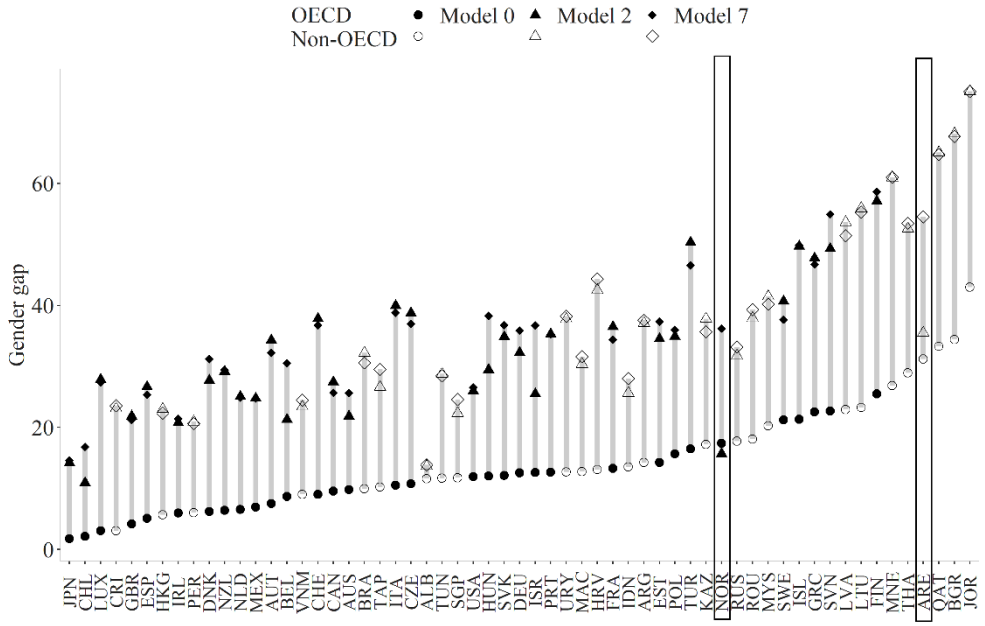
Next, I take a closer look at models M0, M2 and M7 in reading to further examine how the specification of the conditioning model impacts the gender gap. Figure 2.6 hence illustrates the gender gap in reading using model M0 (no conditioning – circle), M2 (just direct individual regressors including gender – 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 half of those, they sit on top of each other. This suggests that, in most countries, the gender gap is not sensitive to the exact specification of the conditioning model (once gender has been included as a direct regressor) with a potential small increase or decrease in the full model.

[12] 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.

There are, nevertheless, some important changes to the results for some individual countries (that are somewhat difficult to explain). Visible differences between M0, M2 and M7 occur in multiple countries. For instance, in Norway and the United Arab Emirates (framed by the two boxes) the estimated gender gap from M2, the model including gender, is even more similar to M0, with a large jump in the magnitude of the gender gap in M7. Looking closer at Norway, gender has a positive latent regression coefficient in both models, but the remaining variables differ significantly. Thereby, principal components (and school IDs) are included in M7 which partly have negative coefficients and also seem to interact with gender. Such changes are perplexing and again suggest that the precise specification of the conditioning model applied can have an impact upon a key aspect of a country's results.

Figure 2.6 Country reading gender gap without conditioning (M0), just with individual direct regressor incl. gender (M2) and with full conditioning (M7)



Note: 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. The two boxes highlight the examples given in the main text for substantial differences between the gender gap for M2 and M7.

Thus far, I 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 is used as 'indirect regressors' – meaning they are only incorporated into the conditioning model having first been pre-processed using a Principal Component Analysis (recall Subsection '2.2.6 How are student background data incorporated into the plausible values?' in '2.2 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 an analysis focusing upon migrant status (one of the most widely used contextual variables from PISA that is an indirect regressor in the conditioning model) are presented in Figure 2.7. These show 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 often sit on top of each other. In other words, for those 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 is -22 points for M3 and -23 points for M7, while it is -20 without any conditioning). Yet there are again some important exceptions to this finding, most notably Norway with a migrant-native reading gap of -43 points under M0, -2 points under M3 and -32 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 10 students are migrants – meaning the fluctuation in the results for this country are unlikely to be due to the small sample size, but I established throughout this chapter that Norway seems especially susceptible to the impact of the conditioning model which can drive such an effect.

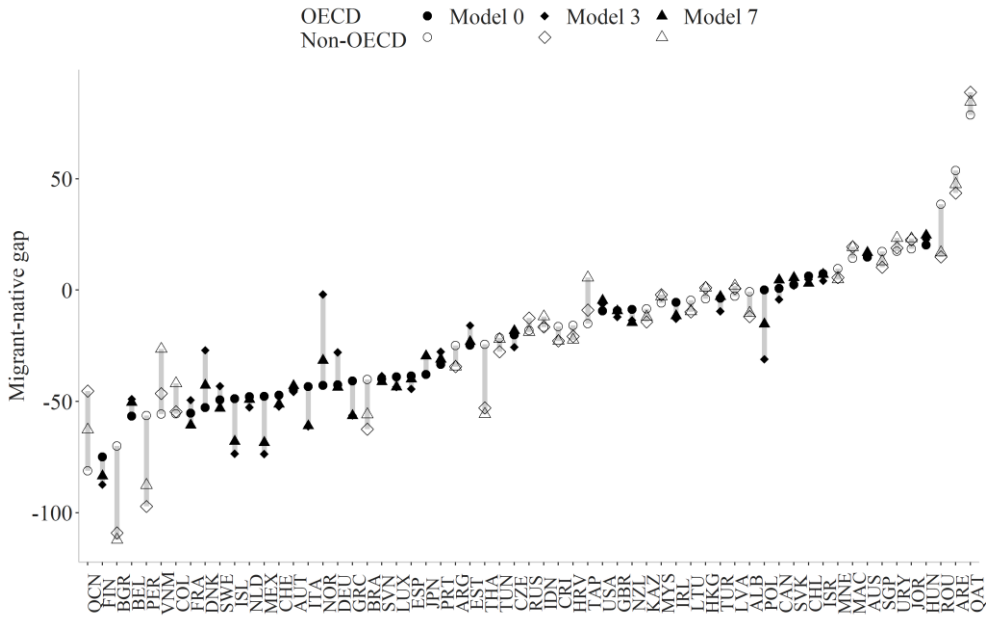
One might be tempted to conclude from this that conditioning does not matter for the gap between migrant and native students in most countries. However, an alternative explanation could be that migration status has not been sufficiently represented within the principal components that form the individual indirect regressors. Would the magnitude of the migrant-native

gaps change if migrant status was included as a direct regressor in the conditioning model instead? I explore this issue in Appendix A.7, where a further alternative version of the conditioning model was computed:

- 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 principal components (PC)s (M7a)

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

Figure 2.7 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)

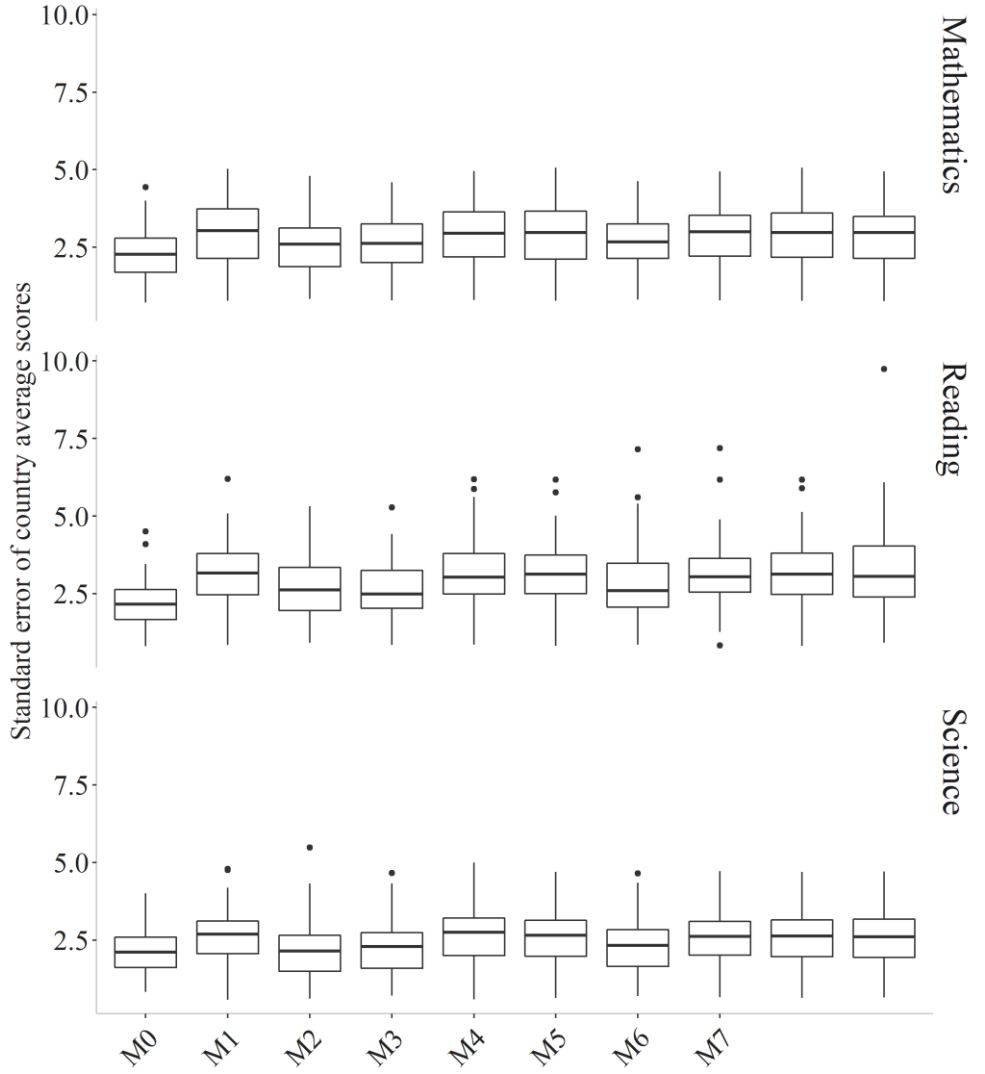


Note: Circles provide estimates without conditioning, diamonds for conditioning only with individual indirect regressors and triangles for full conditioning. Solid markers denote OECD countries and hollow markers non-OECD countries.

2.3.5 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, I therefore investigate how conditioning affects the standard error of country average scores. Figure 2.8 provides a boxplot illustrating how the standard error of the mean changes for different specifications of the conditioning model. One would anticipate that the boxplots should move southwards as one moves from left (M0) to right (M7) – as more information is being used about students to derive the plausible values. But this is not the case; standard errors are typically *higher* once conditioning is used. In fact, in mathematics and reading no country had a smaller standard error when full conditioning was used (compared to no conditioning). In science, only four countries (Singapore, Macao, Estonia, and Canada) experienced an increase in precision when full conditioning was applied. However, in general, no substantial benefit can be found for precision from conditioning, with standard errors actually inflating, if anything.

Figure 2.8 Boxplots of standard errors of country average scores in mathematics, reading and science with different specifications of the conditioning model



Note: 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

2.4 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 underpin 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 scrutinised key aspects of the PISA scaling model. This includes 'conditioning', where background variables are used in the derivation of the PISA plausible values.

This chapter has tried to fill this gap in the literature. Specifically, I 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. I not only document the impact that this has upon average country scores, but also cross-national comparisons of educational inequality (i.e. the spread of achievement) and gaps in performance between different groups (e.g. gender differences).

My results illustrate how the precise specification of the conditioning model does indeed matter, though the impact this has depends upon both the subject and the statistic of interest. In terms of average scores, results for the major domain can be considered 'robust' (i.e. unaffected by whether/how conditioning variables are used). Yet, results for the minor domains are more mixed. Although the specification of the conditioning model has little impact upon cross-country comparisons of average scores in science, the same is not true for reading, where average scores (and, consequently, rankings) change. Rather different results were obtained for educational inequality, where cross-country comparisons in all three domains were sensitive to the specification of the conditioning model. The conditioning model specification was also found to have some impact upon the magnitude of group differences, with particularly big changes observed for gender differences in reading and mathematics in a few countries.

While I 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 the self-computed plausible values and those provided in the OECD PISA database. Although I believe that the approach I have taken provides a sufficient basis for the present study, it is not a perfect replicate for what the OECD (and their contractors) have done. Unfortunately, the OECD do not release their code for how they have constructed the PISA scores (and were unwilling to provide it when requested). To be as open as possible about my own approach (and to allow other researchers to independently scrutinise my findings) I have made freely available the code I have used to produce these results (Zieger, 2021). I now encourage the OECD to improve their transparency, and to do the same.

Second, I focus on the methodology used for one specific PISA cycle (2012). I note that the scaling model (including the conditioning) changed in PISA 2015 and with the introduction of computer adaptive testing in 2018. This means that this chapter is not directly applicable to subsequent PISA cycles, though still yields some important lessons learnt. Finally, I 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 the 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 chapter.

Despite these limitations, I hope this chapter 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 scrutinise the scaling model and its resulting plausible values. The OECD must become more transparent in its methodology and make its technicalities more digestible – particularly to non-specialised audiences. Second, educationalists and policymakers the world over should note from these findings that, while results from the major domains appear to be more trustworthy and robust, measures of inequality

and results for the minor domains in general should be treated with more care. Finally, I question PISA's reliability as a way to compare educational inequality across countries, given the major impact the conditioning model specification can have upon the results.

3 The effect of background variables and design choices on student achievement scores: A simulation study based on PISA 2012

In Chapter 2, I showed that the conditioning model can have a substantial influence on student achievement scores. Yet, the impact was not always easy to explain or anticipate. Furthermore, it was not possible to identify which conditioning model specification captured the underlying 'true' ability best. Using real-life data, this information is not available. In this chapter, I therefore conduct a simulation study based upon the PISA 2012 data and design. This builds upon the foundations laid in Chapter 2. Specifically, as I know the 'true' values used in the data generation process, I can establish which conditioning model specification works 'best' (i.e. was least biased in student achievement scores). I am also able to examine other properties surrounding the conditioning model, such as the impact of alternative test and student background questionnaire designs.

3.1 Introduction

It is hard to imagine educational policymaking without international large-scale assessments (ISLAs), yet the systematic administration of ILSAs in their current form only started in 1995. Since then, studies such as the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) and the International Association for the Evaluation of Educational Achievement (IEA)'s Trends in Mathematics and Science Study (TIMSS) have seen a sharp increase in the numbers of participating countries and educational systems – from 43 in 2000 to 79 in 2018 for PISA and from 41 in 1995 to 64 in 2019 for TIMSS. These studies are regularly used by participating countries as a tool to monitor and evaluate their educational performance and equity (Hopkins et al., 2008), and thus help to shape educational policy as a result (e.g. Breakspear, 2012). This includes the 'PISA shock' in Germany in 2000, which led to major changes in the curriculum (Ertl, 2006), as well as significant impacts in Japan (Takayama, 2008), Denmark (Egelund, 2008) and several other European countries (e. g. Grek, 2009).

Given the prominence and influence of ILSAs, it is of utmost importance that the methodology underpinning them is both sound and transparent. Yet, although the technical reports provide a broad overview of the procedures used (e.g. Martin et al., 2016; OECD, 2014b), the complexity of the approaches means that the processes used to generate PISA and TIMSS test scores is only fully understood by a handful of specialised researchers (Goldstein, 2017). While the broad idea behind the approaches used is explained in general terms (e.g. von Davier et al., 2009), the rationale for certain methodological decisions and specifications (including alterations made to the scaling model over time) are not always clearly discussed and justified in the technical documentation. The comparability and reliability of the results for various different purposes (e.g. whether test scores are really comparable over time) therefore becomes difficult for independent researchers to judge.

As a result, there is an ongoing academic debate about the statistical methodology in ILSAs. While some discuss and challenge the general idea and usage of ILSAs, especially PISA (e.g. Johansson, 2016; Meyer, 2014), there has also been research and criticism of the methodology underpinning these studies even among those who are sympathetic to the approach in general. The studies involve several complex steps in their administration across culturally diverse countries, all of which can potentially affect the results. These include cross-cultural comparability and translation issues, as well as challenges with sampling, non-response and population coverage (e.g. Anders et al., 2021; S. Hopmann et al., 2007; Kankaraš & Moors, 2014; Micklewright et al., 2012; L. Rutkowski & Rutkowski, 2016). Others have questioned key assumptions underpinning the psychometric approach of the studies, such as unidimensionality of the measured latent traits (Wuttke, 2007), or call for fewer but better (non-cognitive) scales, with clearer communication of the complexities ILSAs involve (Avvisati et al., 2019).

Some methodological aspects of ILSAs have naturally been focused upon more in research than others. One area which has received relatively little attention, and is comparably poorly understood, is the way that achievement scores are derived. This includes both the underlying methodology and the

influence of the data choices surrounding it. Most ILSAs employ rotated test designs which means that students only answer a fraction of all questions and, in some studies, not even questions in all domains. Therefore, complex methodology, i.e. 'conditioning', is necessary to derive good achievement estimates. This is where information drawn from the background questionnaires (e.g. gender, parental education) is used to adjust students' estimated achievement distributions, over and above the responses they provided to the test questions. While this technical aspect of studies such as PISA and TIMSS does not often draw great attention, research has shown that it can have an impact upon the results (Mislevy et al., 1992; L. Rutkowski, 2011, 2014).

Indeed, when computing the scores, there are three factors which can have an influence: (a) the form of the data (e.g. are questions asked in all domains and how many?), (b) the quality of the background data (which may be subject to measurement error) and (c) how the conditioning model is specified. With respect to the former, Mislevy et al. (1992) showed that the design of the data (e.g. complete data, data with missing by design) can influence students' scores and that the models need to be chosen accordingly. Regarding the quality of data, Rutkowski (2014) found measurement error in background variables (such as students misreporting parental education) can lead to meaningful under- or over-estimation of group differences. I showed in Chapter 2 that different conditioning model specifications (i.e. which variables the model includes) can have a substantial impact upon cross-national comparisons of educational inequality. Yet, despite this, the technical documentation for PISA, TIMSS and other ILSAs provide little guidance or detail about the conditioning model chosen, and the robustness of results to alternative specifications.

This chapter aims to expand understanding of the conditioning process used in ILSAs and the repercussions of the choices made. As a starting point, I use Chapter 2 which used PISA 2012 data to show that the exact specification of the conditioning models matters. Yet, the reasons and mechanics that drive changes in results remain unclear. This can be challenging in empirical studies, using 'real world' data, where the true value of the latent construct of

interest is unknown, while the data and their structure is fixed. Indeed, in such situations, researchers are restricted to analyse questions which can be naturally answered by existing data. This makes it difficult to establish why non-trivial changes occur and to assess related questions which go beyond the available data.

Simulations, on the other hand, can help to shed light on such matters. When using simulated (rather than real) data, both the 'true' values and those derived from statistical models (plausible values in the case of student achievement in ISLAs) are known. This enables us to not only examine the impact that systematically varying the variables in the conditioning can have, but also change underlying characteristics of the data and judge whether these decisions help or hinder the bias and precision in the estimates of interest. I thus aim to extend the study in Chapter 2 by identifying the factors that lead to changes in the PISA results and study the impact of these choices via a set of simulations.

More specifically, this chapter presents a simulation study drawing cognitive test score and background data following the structure of PISA 2012. I compute eight alternative sets of 'plausible values' (PISA test scores) varying the background variables and data included in their computation, closely following the methodological approach taken in PISA 2012 (and ILSAs more generally). In conjunction with the simulated true values, these computed plausible values allow investigation of: (a) how background variables influence the final PISA scores and whether this varies by country and measure; (b) whether entering background variables directly into the conditioning model (rather than being first combined using principal component analysis) leads to a different result; (c) if less biased population or subpopulation student achievement estimates can be recovered if everyone would have answered questions in all core PISA cognitive domains; and (d) whether substantially different results emerge when using a rotated background questionnaire design instead of a full one.

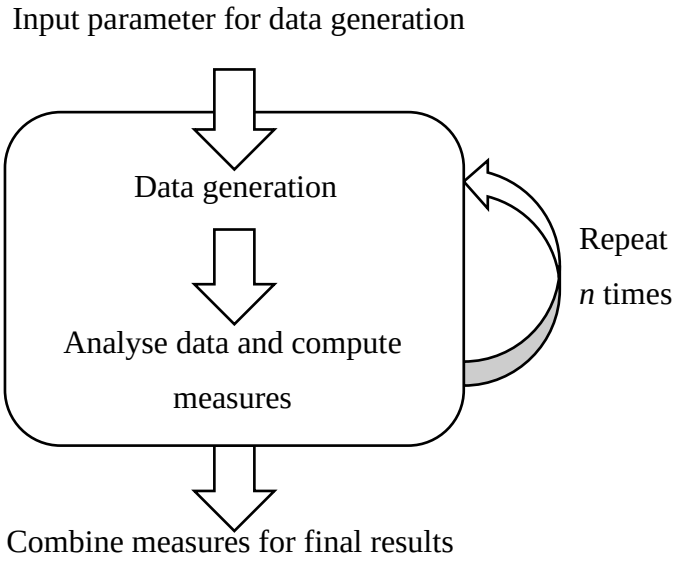
These simulations lead to four key conclusions. First, conditioning model choice matters. Including or excluding a few variables can lead to substantial

changes in bias, especially in PISA 2012's minor domain of reading. There does not seem to be a general pattern of bias which holds true across measures and domains; substantial variation occurs (particularly between measures). Second, while the composition of background variables used in the conditioning model can have a large impact, the same does not hold true for the way the variables are prepared. The difference between entering the variables directly or indirectly (i.e. via first reducing them into a smaller set of principal components) is minimal. Third, when the measures are only based on students who actually answered questions in all domains, the bias in estimates is dramatically reduced in the minor domains. Finally, I find that using a complete (rather than rotated) student background questionnaire generally has no impact on the bias and if it does, the impact is negligibly small. This, in turn, has important implications for the design of ILSAs in the future.

3.2 Methods

The overarching methodological intuition behind this chapter is displayed in Figure 3.1. The simulation (displayed by the box) is a tool which enables the comparison of different models and measures. Thereby, the simulations encompass the 'analysis of interest' (in this case the computation of plausible values using different conditioning models and settings) and account for chance by having *n* repetitions. In each run, data is drawn according to properties defined *a priori*. These data are then used in the analysis of interest. As a consequence, I end up with *n* results from the simulations runs, which are evaluated together with a specific goal in mind. My particular interest is in the bias between the 'true' achievement values (defined within the data generating process) and the plausible values derived from the conditioning model used following the method employed in PISA 2012.

*Figure 3.*1 *Simplified procedure of the simulation in this chapter*



To describe the simulation and subsequent analysis of interest, this section is structured in two parts, based upon Morris et al. (2019). To begin, I provide an overview of the simulation method. This contains information about the data generation process, the estimand and the performance measure. I then move on to describe the 'method inside the simulation', which focuses upon how plausible values are computed and the conditioning model used. All computations within this chapter are done within R (R Core Team, 2019) using the 'lsasim' package for the data generation (Matta et al., 2018) and 'TAM' (Robitzsch et al., 2018) for the computation of plausible values.

3.2.1 Simulation method

This simulation aims to study how certain data characteristics and the inclusion of background variables in ILSA conditioning models affect cross-national comparisons of educational performance and inequality, particularly the bias it may introduce into the results. My particular interest is in estimates of the spread of the achievement distribution (measured as the difference between the 10^{th} and 90^{th} percentile) and gender differences in achievement scores. These are both commonly discussed measures in educational inequality discourse and are the estimands in this analysis. Overall, I look at the following three estimands in this chapter:

- θ_m: Country mean of student achievement,

- θ_g : Gender gap by regressing student achievement on reported gender (indicator for reporting female),
- θ_s : Spread of student achievement by subtracting the 10th percentile from the 90th.

The 'true' value of the estimand (denoted $\theta_l, l \in \{m, g, s\}$) is based upon the simulated 'true' achievement values. In contrast, the estimator of the estimand (denoted $\widehat{\theta_l}, l \in \{m, g, s\}$) is based on the five derived plausible values[13]. I compare eight different versions of plausible values with each other and the true value; therefore, I end up with eight estimators of the estimand $\left(\widehat{\theta_l}^{(0)}, \dots, \widehat{\theta_l}^{(7)}\right)$ and the corresponding estimand (θ_l) in each country and simulation run. The estimators of the different simulation runs of a given conditioning model are evaluated with regard to *bias*, given by the performance measure:

$$Bias_{\theta_l}\left(\widehat{\theta_l}^{(j)}\right) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_{l,i}^{(j)} - \theta_{l,i},$$

where $i = 1, \dots, n_{sim}$ is the number of the simulation run, j the set of plausible values and l the estimand of interest. In the following, estimands that are computed based on the simulated underlying trait are called 'true values' and the ones estimated on the cognitive data are called 'estimated values'. Bias thereby denotes the difference between those two and reflects the deviance caused through the computation model.

The other core part of the simulation method is data generation, which I want to resemble the PISA 2012 data. This data consists of two main parts: (a) responses to test questions and (b) background questionnaire.

PISA tests students in three core domains (mathematics, reading and science) and thereby uses a rotated test design. This means that students in PISA 2012 were randomly assigned to one of 13 booklets. Each booklet consists of four item clusters (groups of items), which are selected from seven mathematics,

[13] In later cycles, PISA uses 10 plausible values instead of five. As a sensitivity check, I redid this analysis with 10 plausible values and found there to be little difference to the substantive conclusions reached.

three reading and three science item clusters. Every year one of three domains (mathematics in 2012) is the 'major domain', meaning that each booklet has at least one mathematics item cluster. Nevertheless, only five of the 13 booklets contain items of all domains. As a result, only around 40% of the students answered questions in all three core domains (OECD, 2014b, pp. 30, 31).

With respect to the background questionnaire(s), only the student questionnaire is administered in all countries. In PISA 2012, this had one common component (including gender, language at home, parental education) for all students and one of three other components randomly assigned to each student (covering topics such as questions about school, problem solving, learning mathematics). Countries could also opt to administer additional questionnaires (e.g. about educational career, information communication technology, information gathered from parents) and/or additional assessments in digital reading, computer-based mathematics, financial literacy and problem solving (OECD, 2014b, pp. 22, 259, 260; see Appendix A.1 for more details).

The 'lsasim' R package (Matta et al., 2018) specialises in the data generation/simulation of large-scale assessments including rotated test designs. I use this package to replicate the design and structure of the PISA 2012 data (OECD, 2014b, pp. 30–32). In doing so, data is drawn from 'known' parametric models in two steps. First, the background questionnaire(s) and the 'true' achievement values are simulated. The known model is thereby constructed with the correlations[14] between the background variables and the plausible values (as proxies for true achievement) as well as the characteristics of the items (mean and standard deviation for numeric items, proportion per category for categorical ones). Second, corresponding cognitive data (item response patterns) is generated. This was done using the published PISA 2012 item difficulties, which were grouped into item clusters and booklets (OECD, 2014b, pp. 406–413). These, in combination with the

[14] The pairwise correlations were extracted from the real PISA 2012 data for each country. Depending on the variable types, the correlations were either Pearson, polychoric or polyserial.

simulated 'true' achievement values, were used to form the parametric model and draw the simulated item response data. The whole data generation process was executed separately for each country. For the sake of simplicity, clarity and run-time, I decided to focus on the eight countries (and their characteristics) listed in Table 3.1 and did not include optionally administered cognitive domains and background questionnaires. These eight countries were chosen because they are the G8 countries with similar economic conditions while displaying clear difference in student achievement. The rotated design of the student background questionnaire was computed 'manually' by randomly assigning students to student questionnaire booklet IDs and setting the corresponding items to missing. Additional non-response was added randomly according to the proportions in the actual PISA 2012 data.

Table 3.1 Sample sizes in PISA 2012 of the countries which are used in the simulation

Country	**Abbr.**	**Sample size**
Canada	CAN	21544
France	FRA	4613
Germany	DEU	5001
Italy	ITA	31073
Japan	JPN	6351
Russia	RUS	5231
United Kingdom	GBR	12659
United States	USA	4978

Note: The sample size shows the number of students with valid data in the major domain, mathematics (OECD, 2014b, p. 241).

The minimum number of simulation runs is defined by:

$$B = \left(\frac{z_{1-\frac{\alpha}{2}}\sigma}{\delta} \right)^2,$$

where δ is the level of accepted accuracy, $z_{1-\frac{\alpha}{2}}$ the $\left(1-\frac{\alpha}{2}\right)$-quantile of the standard normal distribution and σ^2 the variance of the parameter of interest (Burton et al., 2006). A preliminary simulation study (which had exactly the same set-up but was limited to just 100 simulation runs) was used to get first estimates for σ^2. These estimates were then used to determine the number of

simulations within the main study. In this case, the percentile gap in France yielded the highest variance across the simulations with $\sigma^2 = 0.005$. As I want to achieve estimates within 5% significance and 0.005 permissible difference from the true estimand, I set the number of simulation runs to:

$$n_{sim} = 1000 > B \text{ with } B = \left(\frac{z_{0.975}\sqrt{0.005}}{0.005}\right)^2 = 768.29.$$

The number of observations n_{obs} in each simulation run varies by country. For those with sample sizes of less than 10,000 students, the data is drawn according to the original PISA 2012 sample size (see column 3 in Table 3.1). For countries with sample sizes greater than 10,000 students, the number of simulated observations is set to 10,000.[15]

3.2.2 Method inside the simulation

With the help of the method inside the simulation, I want to gauge the effect that choices around the data that are included can have on the plausible values. In order to do so, I compute plausible values in line with common practice in ILSAs. While there are differences between ILSA studies (e.g. PISA 2012 uses a 1-parameter logistic (pl) Item Response Theory (IRT) model versus a 2-/3-pl model in TIMSS 2015; Martin et al., 2016), the general steps in the computation of plausible values remain the same. In this chapter, I use the method used in PISA 2012 (see Chapter 9 and 12, especially pp. 159, 253, 254 of OECD, 2014b) as the starting point.

3.2.2.1 General steps in plausible value computation

Both the students' responses to the test questions and the background questionnaire(s) are used in order to estimate achievement through the following five steps.

- First, an IRT model is estimated on a common sample[16] to determine the item difficulties of the core domains (e.g. mathematics, reading and science in PISA) for all subsequent steps.

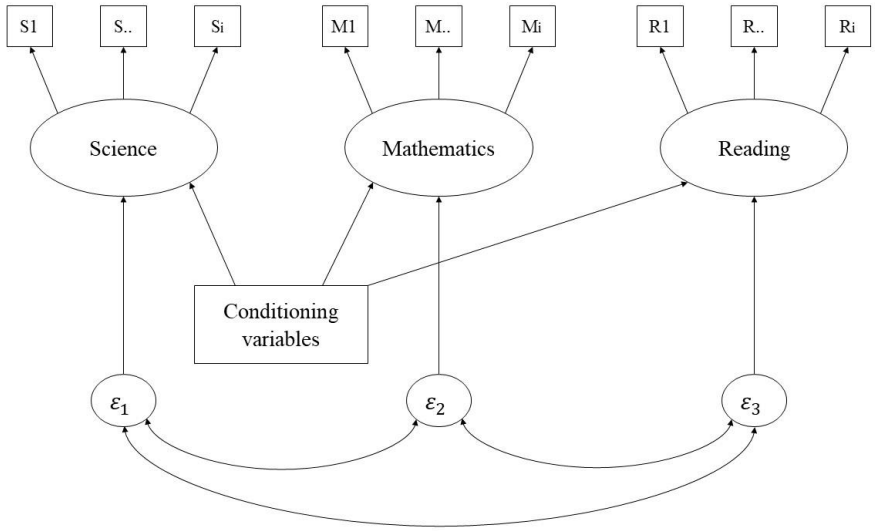
[15] This is done because countries with sample sizes above 10,000 were split into smaller subsamples for the plausible value computation in PISA 2012.
[16] The common sample can consist, for example, of random subsets of all countries or data from all countries which participated in two subsequent cycles.

- Second, background variables are prepared (recoded and/or pre-processed) for use in the plausible value computation. Henceforth, the prepared background variables are denoted ‘conditioning variables’.
- Third, the ‘conditioning model’ estimates the student achievement distributions. In the first step, an IRT model uses students’ responses to the test questions to estimate a preliminary multi-dimensional achievement distribution encompassing the various domains. Subsequently, the conditioning variables are used in a latent regression to adjust the distributions for population characteristics. Figure 3.2 shows a simplified illustration of the conditioning model for PISA.
- Fourth, five plausible values[17] are randomly drawn from each student’s achievement distribution. The concept behind the plausible values can be described as ‘imputations’ for unobserved (latent) student achievement (Mislevy, 1991).
- Finally, via common item equating (or methods with the same aim) these plausible values are transformed onto a common scale with previous PISA cycles. This step aims to facilitate comparisons of student achievement over time (and, hence, carries little importance for this chapter in which I focus on a single PISA cycle alone).

The focus of this chapter is the role of data and background variables in the conditioning detailed in the second and third bullet point above. As a result, I will not address issues relating to the first, fourth and fifth points in this chapter. The most obvious difference in these estimates compared to the official PISA 2012 values is due to Step 5, as PISA scores are transformed onto another scale for cross-cycle comparability.

[17] The choice of five or 10 plausible values in this chapter does not lead to any substantial differences. As five plausible values were used in PISA 2012, I choose that as well.

Figure 3.2 Simplified illustration of the PISA scaling model used to generate the plausible values



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.

3.2.2.2 Background variables in International Large-Scale Assessments

Conditioning models have been used for decades in ILSAs. Yet few people (especially non-specialists) understand how and why background variables are used in the computation of PISA scores, with the technical reports simply describing it as a 'natural extension' of IRT (OECD, 2014b, p. 145).

There is, however, a clear scientific justification for their use. Because of the complex PISA test design (where students only answer a subset of questions within each subject area) conditioning is necessary to reduce *attenuation bias* when comparing test score differences between groups, i.e. it is designed to facilitate unbiased estimations of group differences (Mislevy, 1991; Mislevy et al., 1992). This can be illustrated by a simple example.

Take a hypothetical test where all students are assessed in mathematics, but only half are randomly selected to also receive questions in reading. For those students who answer both reading and mathematics questions, I see that girls outperform boys in reading by 10 points, with there being no gender difference in mathematics. If I were to try to predict reading scores for those

students who did not answer the reading test question, and I did not 'condition upon' gender, girls and boys would be assigned the same reading score (given that boys and girls performed equally well at mathematics). As a result, the gender gap in reading would be estimated as 5 points instead of 10 (i.e. there would be attenuation bias). Therefore, in rotated psychometric test designs where each student only answers a subset of questions (as in ILSAs), population characteristics (such as gender) need to be adjusted in order to produce unbiased results.[18] As in the example above, PISA does not administer questions in all domains to all students. This is different to other ILSAs, such as TIMSS, where every test taker receives questions in all core domains.

The above highlights the importance of the conditioning model in PISA. Nevertheless, it is vital that the model is correctly specified in order to avoid bias (L. Rutkowski, 2014; Wu, 2005). Yet, there is little research or guidelines on the correct specification of conditioning models and the repercussions if they are not implemented appropriately. The documentation of ILSAs also contains little information on the conditioning model selection process or robustness checks, making it difficult to judge this part of the plausible value computation. Furthermore, the same holds true for the robustness of the models depending on whether all students answered questions in all domains or not, or whether the background questionnaire was administered using a rotated design.

3.2.2.3 Plausible value computation in this chapter

In order to investigate how certain choices around the data involved in the plausible value computation influence the PISA results, I simulate data and compute different sets of plausible values. Thereby, I conduct Steps 2–4 (from the bullet-point list above) in each simulation run. The conditioning

[18] Another approach to explain conditioning is in reference to Rubin's (1987) well-known multiple imputation (MI) method. Conditioning models can be seen as MI applied to IRT, treating students' latent abilities as an extreme form of missing data. Following this approach, the same rationale for inclusion of background variables in the estimation of students' latent abilities surfaces: unbiased estimates of group differences (Mislevy, 1991; Mislevy et al., 1992)

model and the draw of plausible values follows the PISA 2012 method and the formulae and annotation used within the OECD technical reports (OECD, 2014b, pp. 144–146), let:

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

Furthermore, the underlying theory of the IRT model and its response vector are adopted from the PISA technical report (for a description and explanation of the IRT model in PISA, see also Chapter 2.2.5 'Replication of the PISA methodology'). The density function of the IRT model without conditioning is defined as:

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

with the density of latent achievement assumed to be normally distributed with $N(\mu, \sigma^2)$.[19]

I want to allow students from different subpopulations to have different abilities. As a result, the density function needs to be altered. The latent variable is, hence, now represented through $\theta_{in} = \boldsymbol{Y_n'\beta} + \varepsilon_n$.[20] Instead of a general mean μ, the abilities of subpopulations are represented through the regression model estimate $\boldsymbol{Y_n'\beta}$. Thereby, $\boldsymbol{Y_n}$ should include all variables which could potentially be of interest for later group comparisons.

The substitution of μ with $\boldsymbol{Y_n'\beta}$ leads to the following adjusted formula:

[19] There are different assumptions which enable the estimation of an IRT model. This 'marginal approach' which specifies the density of latent variables is common in ILSAs but is not the only available one.

[20] ε_n is normally distributed with mean zero.

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

In ILSAs, the student distributions are usually assumed to be multidimensional, including all tested domains (e.g. mathematics, reading and science in PISA). The previous formula can be altered to facilitate the estimation of multidimensional latent variables with conditioning. In this case, $\boldsymbol{\gamma}$ is the multidimensional version of the regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{w_n}$ denotes the equivalent for $\boldsymbol{Y_n}$ and $\boldsymbol{\Sigma}$ is the corresponding variance-covariance matrix for the D dimensions:

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

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.

These formulae set out the theoretical foundation for the simulations. I then implement it empirically as follows.

- The official reported item difficulties from PISA 2012 are used for both the test response simulation and the IRT models.[21]
- The simulated data is used to form eight different settings of involved cognitive and background data for the plausible value computation. Details are described in the next section.
- Using the simulated test data (or subsets), the prepared conditioning variables and the reported item difficulties, the conditioning model is estimated. This is done separately for each country. The models compute student achievement distributions for the three main domains in PISA[22] using a 'divide-and-conquer' approach (Patz & Junker, 1999; van Rijn, 2018). This means that the IRT model and latent regression of the conditioning model are not computed at the same

[21] It is worth noting that I believe that the step difficulty of item PM155Q03D contains a typographical 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 instead of $\tau_1 = -1.569, \tau_2 = 1.569$).

[22] This is a difference to PISA 2012 computation. In PISA 2012, where applicable, additionally administered domains were also integrated into the computation in additional steps.

time, but instead take place sequentially. First, the student achievement distributions are computed via the IRT model [`tam.mml()` from the R package 'TAM' (Robitzsch et al., 2018)]. Second, the distributions are tweaked with the help of latent regression [`tam.latreg()` from 'TAM' (Robitzsch et al., 2018)]. Splitting the estimation in two parts substantially reduces the computational effort and is therefore the default approach in most large-scale assessments (van Rijn, 2018).[23] Quasi-Monte Carlo integration (Pan & Thompson, 2007) with 2000 nodes and convergence criteria of 0.001 for deviance and 0.0001 for the coefficients is used within the computations.

- Finally, I draw five plausible values for each student and domain from the student achievement distributions. Thereby, I draw from an achievement distribution (which is assumed to be multivariate normally distributed) with the help of Monte Carlo estimation with 2000 ability nodes (OECD, 2014b, p. 146).[24]

3.2.2.4 Preparation of different sets of conditioning variables

This chapter highlights the influence and importance of background variables in the plausible value computation in ILSAs. Importantly, these variables are not necessarily used directly within the PISA conditioning model. While all variables from the background questionnaire (in this study, the student background questionnaire, but in PISA also parental, information communication technology and educational career questionnaires, where administered) are included in the conditioning variables ($\boldsymbol{Y_n}$), not all are prepared the same way. Again, I prepared the conditioning variables in line with PISA 2012, with all background variables entering the conditioning

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

[24] Note that I refrain from transforming the values onto a common scale, as is done in PISA. This is because my goal is not to compare PISA scores across cycles, but rather to investigate the effects of conditioning. Estimates are all on the same scale through the item sets used. Further transformation of the scale is therefore not necessary to compare estimates between models.

model as either ‘direct’ or ‘indirect’ regressors (OECD, 2014b, pp. 157, 421–431). These are defined as follows:

- **Direct regressor**. A few selected variables are directly included in $\boldsymbol{Y_n}$ with just some basic recoding. In this chapter, the following variables are treated as direct regressors: gender, grade, mother’s and father’s socio-economic index and booklet IDs.[25] In PISA 2012, school ID, which I did not simulate, was also a direct regressor.
- **Indirect regressor**. The remaining background variables are included in $\boldsymbol{Y_n}$ in a different way in order to reduce dimensionality. First, they are recoded in one or two of the following ways: (a) recoded into indices; (b) dummy-coded if categorical or (c) centred with an additional dummy for missing if numerical. Afterwards, all variables were included in a principal component analysis for mixed data[26] (a combination of principal component analysis and multiple correspondence analysis, Chavent et al., 2014). The indirect regressors consist of as many principal components as it takes to explain 95% of the variance[27]. According to the official documentation, the background variables were not imputed or altered in regard to missing values apart from a missing indicator.

Both direct and indirect regressors are computed separately by country. As a result, the number of regressors varies between countries due to differences in available information (e.g. no variation in grade for Japan) and number of retained principal components.

3.2.2.5 Different settings for the plausible value computation

To gauge the effect that choices surrounding the plausible value computation can have, I use different sets of included cognitive data and conditioning

[25] The contrast coding for booklets was further tweaked so that the information for students who only answered questions in two domains is based on information from all booklets that have items in a domain (OECD, 2014b, p. 157). Additionally, for booklets which included only two domains, the latent regression coefficient is set to 0 in the third domain.

[26] While the usage of a principal component analysis is described in the technical documentation (OECD, 2014b), no details were given, e.g. if a standard principal component analysis is used or one for mixed data. As I have mixed data, I decided to use a principal component analysis specifically for mixed data.

[27] The exact number of retained principal components varies by simulation run and country, but they vary roughly around 100.

variables. The baseline model (V0) does not condition on any background variables and has the same design of major and minor domains and background questionnaire(s) as in PISA 2012 (i.e. not all students answer questions in all domains and the background questionnaire has a rotated design). Five different conditioning model specifications (V1–V5) are estimated with this PISA 2012 design but including different combinations and alterations of direct and indirect regressors into the conditioning model. These are complemented by another two settings (V6, V7) which use a fixed set of conditioning variables but are based on slightly alternative versions of the data. In the end, I compute eight different sets of plausible values in each simulation run, based upon eight different settings. These versions are denoted as follows in Table 3.2:

Table 3.2 Definition of the eight different settings for the plausible value computation

Model	Direct regressors	Indirect regressors	Variable preparation	Questions in all three domains	Background questionnaire
V0	-	-	-	No	Rotated
V1	X	-	Direct	No	Rotated
V2	-	X	Indirect	No	Rotated
V3	X	X	Direct/indirect	No	Rotated
V4	X	X	All direct	No	Rotated
V5	X	X	All indirect	No	Rotated
V6	X	X	Direct/indirect	Yes	Rotated
V7	X	X	Direct/indirect	No	Full

Note: X in the column 'Direct regressors' denotes that the direct regressor variables from PISA 2012 (gender, grade, mother's and father's socio-economic index and booklet IDs) are included in the conditioning model. X in the column 'Indirect regressors' denotes that the remaining background variables are included in the conditioning model. The variable preparation corresponds to Subchapter 3.2.2.4; 'Direct/indirect' denotes that the conditioning variables are prepared as in PISA (part as direct and part as indirect regressors). 'No' in column 'Questions in all three domains' means that all students are used, whereas 'Yes' only considers students which were administered questions in all three domains. The column 'Background questionnaire' denotes whether a rotated or full background questionnaire is used.

The first four models (V0–V3) are used to evaluate how conditioning variables affect the achievement scores and whether this varies between countries and measures. Models V3–V5 help show whether there is a difference depending on how the variables are prepared (i.e. whether they are included in a PCA first or not). When I look at differences between major and minor domains, I focus upon model V3, which is closest to the conditioning variables used in PISA 2012, but vary the set of students included in the computation in model V6. Finally, the difference between model V3 and model V7 shows the difference that stems from using a full background questionnaire versus a rotated background questionnaire design.

3.3 Analyses

3.3.1 How does conditioning affect achievement measures?

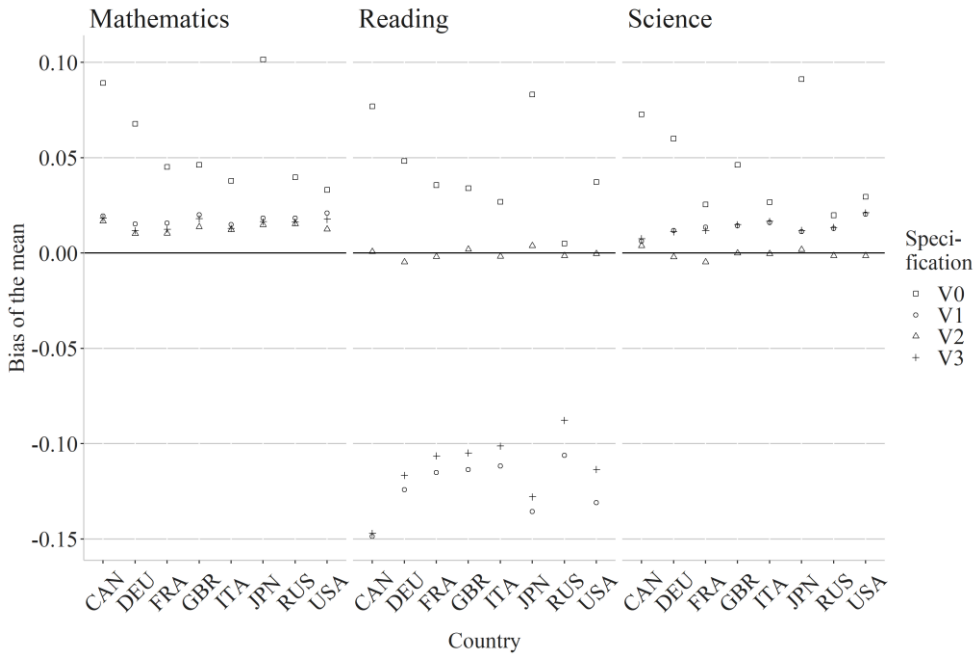
Figure 3.3 shows the bias of the mean (difference between the estimated and 'true' achievement) in standard deviations for four different conditioning model specifications. The horizontal line highlights the optimal case where no bias is present. V0 (square – no conditioning) acts as the baseline model and V3 (cross – all regressors) as the full model, where V1 (circle – direct regressors only) and V2 (triangle – indirect regressors only) are variations.

In mathematics and science, the models with conditioning (V1–V3) outperform the baseline (V0), in terms of minimising bias in the mean score. Yet, there are still differences between the various conditioning model specifications. This is more notable in science, where the average bias across countries is 0.046 standard deviations when no conditioning is used, compared to 0.013 standard deviations in for V1, -0.001 for V2 and 0.014 for V3. For mathematics, the bias amounts to 0.058 for V0 and between 0.013 and 0.018 for the remaining models. While there is a clear decline in bias when conditioning is used, the absolute magnitude of the bias in both science and mathematics is small.

The results for reading give a different picture. While the bias of V0 and V2 behave similarly to mathematics and science, it increases significantly for V1 and V3. This suggests that one or more of the direct regressors has a

substantive undesirable influence, introducing significant bias into the estimated plausible values. The bias increases from 0.043 in V0 and 0.000 in V2 to -0.123 in V1 and -0.113 in V3. This means that, depending on the choice of the specification of the conditioning model, country averages can be biased by more than 0.1 standard deviations (equivalent to 10 PISA points).

Figure 3.3 Average bias of the country means in standard deviations



Note: V0: No conditioning variables, V1: Individual direct regressors only, V2: Indirect regressors only, V3: Individual direct regressors and indirect regressors.

As reading yielded some unexpected results with bias increasing as soon as direct regressors were used, I decided to conduct some further investigations into direct regressors. As a result, I ran four sub models of the conditioning model with direct regressors only (V1.a: test booklet ID only, V1.b: gender only, V1.c: socio-economic index only, V1.d: grade enrolled). This revealed that the inclusion of booklet ID in the latent regression introduces substantial bias to the country averages. Further investigations showed that removing booklet ID from the conditioning model led to the least biased result in most cases. Including booklet ID in the latent regression without any restriction was counterproductive (in PISA 2012 the latent regression coefficients for

booklet IDs were set to zero for the domains which the booklet did not include) and led to some substantial increase in bias in some cases, especially in reading. See Appendix B for more detail.

Recall that the rationale for conditioning is to yield unbiased estimates of group differences, such as the gender gap in achievement. Here, the simulated data is used to estimate the gender gap under different conditioning model specifications by regressing gender (an indicator for girls) upon the simulated PISA plausible values. In this project, the 'true' gender gap (on average, across countries) is equal to 0.059 in mathematics, -0.181 in reading and 0.012 in science. Table 3.3 displays the bias between the 'true' and 'estimated' gender gap across countries. In line with multiple imputation theory, all biases have the opposite sign of the initial gender gap, meaning that the estimates are attenuated. As expected, the bias in the gender gap shrinks close to 0, especially in mathematics and science, as soon as gender is introduced into the conditioning model, meaning that attenuation is significantly reduced. The conditioning model specification without gender, V2 (indirect regressors only), on the other hand, still displays substantial amounts of bias, especially in reading.

Table 3.3 Average bias of the gender gaps across countries

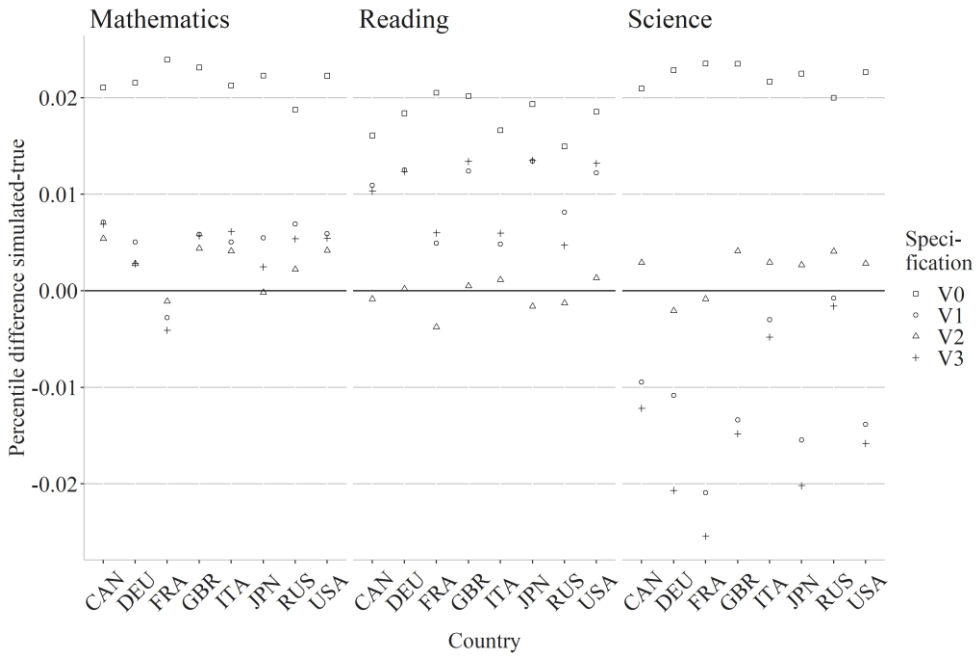
	Average bias of the gender gap		
Model	**Mathematics**	**Reading**	**Science**
V0	-0.056	0.153	-0.019
V1	-0.001	0.007	-0.002
V2	-0.031	0.093	-0.023
V3	-0.001	0.008	-0.001

Note: V0: No conditioning variables, V1: Individual direct regressors only, V2: Indirect regressors only, V3: Individual direct regressors and indirect regressors. Bias is reported in standard deviations on average across the countries.

The difference between the 90th (P90) and 10th (P10) percentile, i.e. the spread of achievement scores in a country, is another way to look at inequality. Figure 3.4 shows how this measure is affected by different specifications of the conditioning model. While there are systematic differences between the model specifications in all domains, the magnitude of the bias for this

measure is small in all cases. Indeed, even at the extreme, the average bias is still below |0.03| of a standard deviation. When no conditioning is used, the bias in P90–P10 is around 0.02 standard deviations (independent of the domain) – with the simulated percentile gap consistently being slightly bigger than the 'true' value. In mathematics and reading, the bias in percentile gaps drops consistently as soon as conditioning is applied. Thereby, the values for V2 (with indirect regressors only) show the lowest amount of bias and tend to be around zero. V2 also produces the least biased results for science on average. There, there is a systematic shift in the bias – including a change of sign – when direct regressors are included in the conditioning model. Like the country averages in reading, this change is triggered by the booklet ID in the direct regressors (see Appendix B for more information). Nevertheless, bias remains comparatively small for all specifications.

Figure 3.4 Average bias of the 90th–10th percentile difference in standard deviations



Note: V0: No conditioning variables, V1: Individual direct regressors only, V2: Indirect regressors only, V3: Individual direct regressors and indirect regressors. Bias is shown in standard deviations for each country.

The previous two graphs combined with the last table highlight how including or excluding variables in the conditioning model can affect bias. Conditioning

model specification consistently affected the amount of bias, but not in a uniform way across measures and domains. In most cases, accuracy increased but that does not hold true for all cases. In some, bias increased after adding variables, e.g. director regressors and country means in reading or spread in science. Additional analyses showed that indeed one block of variables (booklet IDs) can introduce substantial bias to the results. More generally, there are some similarities between the major domain, mathematics, and the minor domains, reading and science, in terms of the behaviour and magnitude, but this depends largely on the measure, e.g. mathematics and science show similar patterns for country averages, but differences for spread, where mathematics and reading display similar findings. In general, the picture can be quite different between measures. With regard to countries, there does not seem to be a clear ranking in terms of bias across measures. Some countries with comparatively low bias in one measure can have larger bias in comparison in other measure.

3.3.2 What is the effect of entering variables directly or indirectly into the conditioning model?

One step of the score computation which changes across the PISA cycles is how the conditioning variables are used. PISA used a mixture of direct and indirect variables (as explained in the methods section) from 2000 until 2012. However, from 2015 onwards, all background variables entered as indirect variables.[28] As a result, I want to investigate whether these different approaches lead to the same results. The full model (V3; including both direct and indirect regressors) was, therefore, re-estimated in two further ways: (a) all variables treated as direct regressors (V4) and (b) all variables treated as indirect regressors (V5). In V3 and V4, booklet IDs are treated as direct regressors and some restrictions are applied to the coefficients in the latent regression (i.e. the coefficient of booklet IDs which only included two domains is set to zero for the third uncovered domain). In V5, booklet IDs are

[28] But in this case only retaining the sufficient number of principal components for 80% of the variance explained instead of 95% of variance explained, as in PISA 2012. Furthermore, the maximal number of principal components was limited to 5% of the raw country sample size.

included as indirect regressors. Thereby, they cannot be 'untangled' from other variables in the principal components and no restrictions in the latent regression can be applied. To enable a fair comparison between direct and indirect regressors (V3) and all variables as indirect regressors (V5), I compared V5 to a tweaked version V3 (denoted V3a), which has no restrictions for booklet IDs in the latent regression.

Table 3.4 shows the average bias across countries for the three different measures and domains. For all domains and measures, the effect of the variable preparation is minimal, with differences of less than 0.01 standard deviations (i.e. approximately 1 PISA test point) between V3 and V4, as well as between V3.a and V5. Furthermore, there is no clear pattern emerging of one of the approaches being any better (i.e. less biased) than the other. Overall, the preparation of the conditioning variables leads to very small differences in terms of magnitude and has just a fraction of the effects that other conditioning model choices have.

Table 3.4 Comparison of the bias of country averages, gender gaps and 90th–10th percentile differences when using different forms of conditioning variable preparation

	Mean	**Gender gap**	**P90–P10**
Mathematics			
V3	0.015	-0.001	0.004
V4	0.015	-0.002	0.005
V3.a	0.009	-0.016	-0.027
V5	0.010	-0.016	-0.025
Reading			
V3	-0.113	0.008	0.010
V4	-0.112	0.009	0.010
V3.a	-0.337	0.073	0.033
V5	-0.331	0.076	0.031
Science			
V3	0.014	-0.001	-0.014
V4	0.013	-0.001	-0.014
V3.a	-0.012	-0.011	-0.066
V5	-0.006	-0.011	-0.064

Note: V3: Individual direct regressors and indirect regressors, V3.a: Individual direct regressors and indirect regressors without regression coefficient restrictions for booklet IDs, V4: All variables treated as direct regressors, V5: All variables treated as indirect regressors. Bias is shown in standard deviation on average across countries.

3.3.3 Would the scores change if all students answered questions in all domains?

As explained above, less than half of the students in PISA answer questions in all domains. In the minor domains, where not all participants answered questions, the bias tends to be larger. This is especially true for reading, where more bias in the scores is found than in mathematics. I am consequently interested in whether the bias is smaller when focusing upon students that answered questions in all three of the core PISA subjects. To do this, I re-estimated the full model (V3) using only the subset of students that answered at least one question in each domain. Achievement levels should thereby not vary between groups, as students are randomly assigned to booklets. Yet, even if there is variation between subsets of students, I am not examining the absolute achievement level of the students, but the bias – the difference between the true and estimated value.

Table 3.5 highlights that bias can decrease substantially if all students answer questions in all domains (rather than the situation in PISA 2012, where some students did not answer questions in some domains). As expected, this leads to improvement in terms of bias for both minor domains, reading and science. Especially for country averages in reading, which are heavily biased through the inclusion of booklet IDs in the conditioning model, this helped to counteract the bias – it dropped by roughly 60% or 0.075 standard deviations (equivalent to more than 7 PISA points). Similar or higher percentage improvements were also found for country averages in science and percentile differences in reading. Overall, administering questions in all domains to everyone had a positive influence on all results in the minor domains. The results in the major domain, mathematics, remained relatively stable with the potential for a slight increase in bias (most likely due to achievement scores now being based on only a subset of booklets).

Table 3.5 Comparison of the bias country averages, gender gaps and 90th–10th percentile differences in reading and science based on the students who answered questions in all three core domains

	Mean	Gender gap	P90–P10
Mathematics			
V3	0.015	-0.001	0.004
V6	0.020	-0.002	0.002
Reading			
V3	-0.113	0.008	0.010
V6	-0.047	0.005	0.003
Science			
V3	0.014	-0.001	-0.014
V6	0.003	-0.001	-0.010

Note: V3: Individual direct regressors and indirect regressors, V6: Individual direct regressors and indirect regressors, but only based on students which answered questions in all three domains. Bias is shown in standard deviations on average across countries.

3.3.4 What is the impact of the student background questionnaire design on the plausible values?

In contrast with the other cycles of PISA, not all questions of the compulsory student background questionnaire were administered to all students in PISA 2012. Thereby, all students saw one core component of the questionnaire (including gender and parental socio-economic status) but only a fraction of the remaining questions. As a result, the indirect regressors were based on questions with high proportions of missing. I am interested to see whether it would make a difference to the plausible values if all questions were administered to all students. In order to do this, I skip the step in the simulation data generation where certain questions in the background questionnaire are set to missing according to their student questionnaire booklet ID. This way, I end up with data as they would have been if all students were administered all questions. Model V7 has the same set of conditioning variables as V3 but is based on the 'full' background questionnaire data.

Table 3.6 shows the difference it would make if all students had been administered all student background questionnaire questions (V7). Overall, there is little to no difference in all domains. Bias stays the same in most cases

and there was a difference of |0.003| standard deviations – less than a third of PISA point – in extreme cases. It can therefore be described that using a rotated background questionnaire (at least in the form of PISA 2012) has minimal to no impact on country averages, gender gaps and percentile differences. The usage of rotated background questionnaire has been discussed in academia before. While von Davier (2013) dissuades the usage of rotated background questionnaires based on theoretical deliberations, Rutkowski (2017) neutrally explains the benefits and limitations of different questionnaire designs and the advantages of using a non-full design. In line with my results, a retrospective simulation of a rotated background questionnaire, similar to the one used in PISA 2012, also found negligible impact on student achievement measures (Adams et al., 2013). As a result, analyses based on real-life data support the benefits of rotated background questionnaires. This is especially true in comparison to the repercussions of other choices regarding the conditioning model and used data. It in turn has important implications for designs of ILSAs in the future, suggesting that more background data can be collected about pupils (via the use of rotated questionnaires) without biasing the student achievement results.

Table 3.6 Comparison of the bias of country averages, gender gaps and 90th–10th percentile differences in reading and science if all students were administered all questions in the student background questionnaire

	Mean	**Gender gap**	**P90–P10**
Mathematics			
V3	0.015	-0.001	0.004
V7	0.015	-0.002	0.004
Reading			
V3	-0.113	0.008	0.010
V7	-0.110	0.008	0.010
Science			
V3	0.014	-0.001	-0.014
V7	0.013	-0.001	-0.014

Note: V3: Individual direct regressors and indirect regressors, V6: Individual direct regressors and indirect regressors, but based on a background questionnaire which was simulated as complete. Bias is shown in standard deviations on average across countries.

3.4 Conclusion & discussion

Over time, ILSAs have gained in both size and importance, and are now widely used in education policymaking. Yet, despite this influence, the methodology underpinning these studies and the associated analytical decisions remains poorly understood by most stakeholders – including researchers – who use the data and the results. This is especially true with respect to the role that student background data plays in the derivation of the 'plausible values'.

This chapter aims to broaden understanding of this topic. In particular, I seek to shed new light on how conditioning variables and data choices influence the results of ILSA studies, such as PISA. This has been done by simulating data in the style of ILSAs, specifically PISA 2012, and computing different sets of plausible values based upon different conditioning models and data situations. These are subsequently evaluated with respect to the bias that they introduce. Thereby, I assess two main areas of interest: (a) the influence of background variables on the scores and (b) the impact of test and questionnaire design decisions. I start by systematically changing the variables which are used in the conditioning model and assessing how they influence the plausible values. I also consider whether there are differences across different countries and subject domains. This part does not only consider the impact of making different variable selections, but also how these variables are pre-processed and prepared. Specifically, I compare results from conditioning models where variables are entered directly to those where they are pre-processed using a principal component analysis (as is commonly applied in ILSAs such as PISA). Secondly, in contrast to other ILSAs (such as TIMSS), PISA does not administer test questions in all domains to more than half of the students. I am interested to see how this data choice affects the results: I investigate whether bias in the estimates of student achievement could be substantially decreased if all students answered questions in all domains. Furthermore, one specific trait of PISA 2012 is that the student background questionnaire also had a rotated design, meaning not all students were administered all background questions. As a consequence, I

also examine the previous issue for the student questionnaire to see whether using a rotated design there also affects the results.

This simulation confirms that conditioning model choices matter. It shows that even making subtle changes – such as adding or removing a few selected variables – can have a non-trivial impact upon the results. The magnitude hence strongly depends both on the domain and the measure. While the impact tended to be larger in the minor domains, where fewer children have completed test items, and could be substantial, in some cases the impact was small to negligible. Furthermore, I could find no uniform pattern of bias across countries and measures. While I find that conditioning model choice matters, the same does not hold true with respect to variable preparation (i.e. whether covariates enter the conditioning model as direct or indirect regressors). Hence it does not seem to matter how the selected variables are pre-processed; what matters more is the choice of the variables. Furthermore, this simulation suggests that bias in the minor domains, which is introduced through the conditioning model specification, can be effectively counteracted and reduced by up to roughly 75% if all students answer some questions in all domains. Finally, I found that the same does not necessarily hold true for the design of student background questionnaires. There was essentially no difference in bias between the PISA background questionnaire design and a full design in most cases.

These findings have important implications. One key result is that it makes a substantial difference in terms of bias whether all students answer questions in domains in which they are scored. This is especially true if a model is misspecified, which can seldomly be determined with real-life data. While asking every student questions in all domains is the standard in multiple ILSAs, it is not in all – such as PISA. Furthermore, having students answer questions in all domains turned out to be substantially more important than having a complete background questionnaire design. As ILSAs usually use complete background questionnaires (PISA 2012 was an exception), one idea could be to change the form of the background questionnaire and thus shorten it in favour of more cognitive items. Another possibility regarding the student background questionnaire could be to use an incomplete design and add

additional topics to get more contextual data whilst keeping it the same length for the individual student. With respect to the conditioning model, this research showed model specification matters and can impact country scores. Yet, the technical reports fail to offer rationale, details, and guidance on analytical decisions behind it. As small decisions behind these models can make a difference, this lack of information makes it difficult to judge and replicate the details. The publication of more details and sensitivity analyses as well as an open scientific discourse about the repercussion of those models should be encouraged.

I believe that these simulations highlight some important properties about the conditioning applied in international large-scale student assessments. Yet this study is not without its limitations and should be interpreted accordingly. One caveat is that each ILSA uses a slightly different variant of the test design. I have focused upon PISA 2012 due to it being part of a well-known ILSA series and having a particularly interesting set of design characteristics. Nevertheless, the topics investigated in this study are, to some extent, common amongst all ILSAs, with the results providing at least indicative evidence outside of this single study alone. A second caveat is that I have tried to replicate the PISA 2012 method – as detailed in the technical report – as closely as possible. While I believe I have achieved this sufficiently to meet my study aims, I cannot completely rule out there being some differences due to a lack of clarity about certain aspects of the PISA 2012 method within the technical reports. Relatedly, it has also not been feasible for me to simulate student data in organisational units (e.g. schools or classrooms) and to model the relationship between these and student characteristics.

4 Group comparisons in PISA: What can go wrong along the way? A case study of differences in achievement by parental education in Germany

The previous two chapters highlight how student achievement scores can be impacted through the conditioning model and related decisions and properties. Yet, both studies are very focused upon a single aspect of ILSAs – the conditioning model – and directly involved properties, such as the data preparation and design. Looking more broadly, ILSAs are large, complex endeavours that require and involve many more properties and decisions. None of these are completely separate from all others; they are all intertwined. In this chapter, I expand my focus and take a more comprehensive perspective of the process behind measuring student achievement and potential sources of bias in ILSAs. To do so, I make use of the total survey error framework to identify six statistical and psychometric properties, which can potentially introduce bias. I closely scrutinise their impact on group comparisons using a case study.

4.1 Introduction

International large-scale assessments (ILSAs), such as the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) and the International Association for the Evaluation of Educational Achievement (IEA)'s Trends in Mathematics and Science Study (TIMSS), are surveys which aim to collect information about the cognitive skills of students around the world. For more than 20 years they had considerable influence on both education policy and daily school life. As ILSAs provide student achievement scores that are considered to be comparable across countries and over time, they have long played an important part in national and international education debates and in subsequent policy decisions (Breakspear, 2012; Hopkins et al., 2008).

These studies do not only gather data about students' cognitive skills, but also administer extensive background questionnaires, which include questions about their socio-economic background, upbringing, and attitudes. As a

result, ILSAs, such as PISA, are increasingly being used to analyse the association between different background variables and student achievement across countries and over time (e.g. Davoli & Entorf, 2018 for policy application; Jerrim, 2013 for research; OECD, 2019c for reports of ILSAs). Among other findings, stark differences in educational outcomes by socio-economic status were found; a topic which has become a key area of public policy interest and intervention (e.g. Carpenter et al., 2013; Kaiser, 2018; Karsten, 2006; Volante et al., 2019). The influence of parental background and its transmission has been widely researched by academics spanning across different disciplines (most notably sociologists and economists) and shown to affect children's educational achievement (e.g. Davis-Kean, 2005; Guryan et al., 2008; Ludeke et al., 2021; Pishghadam & Zabihi, 2011; Yeung et al., 2002). On the other hand, research has also shown that education can act as an important mediator between social origin and destination (e.g. Breen & Jonsson, 2005; Jerrim & Macmillan, 2015), arguing for a need to level the education playing field. In an ideal world, primary and secondary education should at least provide everyone with a set of core basic abilities that equip them with the skills needed for later life.

Of course, if this ideal is to be implemented, these associations – including the role of education – need to be understood in detail first. A common way to achieve this is studying patterns in quantitative survey data, such as ILSAs. However, when doing so, one needs to be aware of the limitations with such data. Specifically, it is important that the methodology and measures underpinning the data are sound and valid. Although this may sound simple, the reality is not quite so straightforward. ILSAs are large, complex projects that require great amounts of work, organisation, and effort, gathering data from hundreds of thousands of children from many countries. While numerous steps are taken to ensure data quality, there remain ways in which bias can creep in. As early as the 1940s, research started to engage with the 'total survey error' (TSE) framework, which aims to evaluate the 'usefulness' and 'meaningfulness' of surveys in general by comprehensively considering various sources of error and bias (Groves & Lyberg, 2010). Data quality can be affected through numerous pathways. This can happen if the data does not

represent the underlying population well due to factors such as survey non-response, missing data (amongst respondents) and coverage, as well as sampling error. Yet, even if the sample covers the population perfectly, error can still be present in the data due to the measurement itself. Bias can be introduced through the actions of respondents (e.g. inaccurate reporting, misunderstanding), the mode of the study, the instrument or even (data) processing post-survey. All of these aspects and pathways can affect the distribution of the quantity of interest, which in turn could lead to substantial under- or over-estimations of socio-economic group comparisons (Billiet & Matsuo, 2012).

In this chapter, I want to highlight different sources of possible bias that should be considered when conducting analyses using ILSA data. Thereby, drawing upon the TSE framework, I investigate six factors that could potentially bias socio-economic comparisons of educational achievement: (i) Schools and students are sampled in ILSAs, but some schools refuse to participate, and students do not show up to the test. There are also important issues due to (ii) non-coverage (Anders et al., 2021) and (iii) missing item data for crucial information, such as parental education (L. Rutkowski, 2011). As a result, ILSAs such as PISA continue – as all surveys – to have issues with missing data (L. Rutkowski, 2011; Wise, 2009). Furthermore, (iv) some students may also misreport information about their family background, particularly as they often act as proxy respondents for their parents (e.g. children report their parents' level of education or occupation; Jerrim & Micklewright, 2014). Moreover, (v) due to the way that achievement scores are constructed in ILSAs, students' background variables can affect their cognitive scores (von Davier et al., 2009), leading to potentially biased estimation of socio-economic group differences (L. Rutkowski, 2014). Finally, (vi) it should be verified whether socio-economic background measures (such as parental education, captured within the International Standard Classification of Education (ISCED) framework) validly measure the construct in the international context and whether the students' responses are correctly coded according to it (D. Rutkowski & Rutkowski, 2013).

The key contribution of this chapter is to add to the existing literature by undertaking a comprehensive review of possible errors – from sampling to missing data and computing cognitive scores – that may affect socio-economic group comparisons in international studies. Previous research has investigated several aspects of ILSAs which can introduce error (e.g. Anders et al., 2021; Heine et al., 2017; D. Rutkowski & Rutkowski, 2013; L. Rutkowski, 2011). Yet, few studies have considered multiple aspects at once or provided a comprehensive review. I address this aim by conducting a case study for highest parental education in Germany for PISA 2012. This particular choice allows us to investigate the six aspects mentioned above in-depth, as I focus on a feasible setting with a particularly rich data situation. A subsection is devoted to each of these six aspects, where I gauge the impact that each has on socio-economic achievement gaps (as measured by differences in PISA scores by level of parental education). I thus aim to foster awareness and understanding for potential sources of bias in group comparisons and the necessity of careful consideration of data quality and measure validity – over and above the specific setting of this case study.

My analyses show that some – but not all – of the six aspects considered impacted the results in a substantial way. Survey and item non-response, (correct) coding of a parental education measure and the quality of students as proxies for their parents' background were all found to be particularly important. Survey and item non-response are found to be related to both highest parental education and student achievement, meaning group comparisons are not representative of the wider population. When checking the coding of the German items to the ISCED scale, I found a procedural error, leading to a substantial change to the distribution of parental education. Moreover, I found that only around half of German students provided the same response about parental education level as their parents – with the socio-economic achievement gap changing depending upon whose response is used. Overall, I find that the data on highest parental education in Germany suffers from serious flaws. These results highlight that the background variables should be checked carefully and not be trusted blindly.

This chapter proceeds as follows. Section 2 elaborates on the total survey error framework and sources of error investigated in this chapter. Section 3 describes the data I use and the methodology, during which I also explain why I use Germany's PISA 2012 data as a case study. Section 4 presents results where I sequentially focus upon six sources of error in socio-economic group comparisons. The chapter is then wrapped up in Section 5 with conclusions, recommendations, and limitations.

4.2 Total survey error and socio-economic group comparisons

This chapter is situated in the total survey error framework literature in ILSAs, in order to assess the quality and meaningfulness of PISA data and group comparisons (Schnepf, 2018). The total survey error has been defined in different ways by different authors, with multiple typologies and schemas now existing (Groves & Lyberg, 2010). In this chapter I use the following description from Groves and Lyberg (2010, p. 850) as a guideline for my analysis:

> *Inherent in the term total survey error is attention to the entire set of survey design components that identify the population, describe the sample, access responding units among the sample, operationalize constructs that are the target of the measurement, obtain responses to the measurements, and summarize the data for estimating some stated population parameter.*

In accordance with this description, I identify six issues that may affect the validity of socio-economic gaps in educational achievement using ILSA data. These are: (i) identifying the target population; (ii) survey non-response; (iii) item non-response; (iv) socio-economic group measurement; (v) measurement error in indicators of socio-economic background; (vi) construction of ILSA test scores. I do not claim that this completely captures all aspects of total survey error. But rather, I provide a comprehensive review of (potential) sources of bias and illustrate the importance and impact of addressing these in common analyses with such data.

4.2.1 Identifying the (target) population

Coverage addresses the degree to which students from the population are covered and participating in an ILSA. This is influenced by multiple factors. First, the target population must be defined. Although this may seem trivial at first, it can have an important impact upon estimates of socio-economic inequalities. For instance, differing school enrolment rates between countries can lead to results which are counter-intuitive and easy to misinterpret (Education Datalab, 2017). Second, the whole target population may not be covered in a survey. It is common that some groups of the target population are not covered due to pre-defined reasons such as accessibility or feasibility. For example, in some countries children with special educational needs (SEN) are excluded from the study, which will have an impact upon estimation of socio-economic gradients if it means certain groups (e.g. children from lower socioeconomic status (SES) backgrounds) are more likely to be excluded. This becomes even more complex in cross-national comparative research, as there are general guidelines about the definition and treatment of SEN children, but the implementation is up to each country. As a result, exclusion and inclusion rates differ drastically between countries (LeRoy et al., 2019).

4.2.2 Survey (unit) non-response

All social surveys are subject to some degree of non-response. In large, cross-national studies of school children, non-response can occur at the school level (as the primary sampling unit) and at the student level. Such non-response is likely to be driven by conscious decisions made by schools and students to not participate in the study but could also be due to technical problems with test or survey administration. If such non-response occurred completely at random, socio-economic group comparisons would not be biased. Yet, in reality, such a strong assumption is highly unlikely to hold. For instance, Heine et al. (2017) showed via an extension to PISA in Germany that non-participation at both the student and school level was related to family background and socio-demographic measures, as well as prior attainment.

When schools and students selectively choose to not respond, this can lead to substantial bias in ILSAs results (Anders et al., 2021).

4.2.3 Item non-response

In contrast to unit non-response, item non-response refers to survey units (e.g. students, parents or schools) that participated in the study, but who did not answer a particular question (or set of questions). This can be done intentionally, by skipping a question, or unintentionally because they were either out of time or the item was not administered. While missing responses are unproblematic when the data is missing completely at random (such as missing due to test design), it can bias results if survey units with particular characteristics choose not to answer specific questions (Rubin, 1976). Previous research has found this to be a problem for measures of socio-economic status in large-scale international assessments. For instance, Caro & Cortés (2012) found that students missing socio-economic status data differed systematically from those with complete data. There may be bias in estimates of socio-economic gaps as a result of using such data.

4.2.4 Constructing and operationalising a comparable group measure

In order to reliably compare socio-economic differences in outcomes across countries, it is vital that the primary measure of interest (family background) carries the same meaning in each nation. Yet, in a diverse world few things have the same meaning everywhere. Take, for example, the case of education. Each country individually decides the organisation and content of its education system, resulting in different qualifications and knowledge at graduation. This, in turn, poses a major challenge for cross-national comparisons of educational qualifications, both amongst parents and children. As a result, considerable time and effort has been invested into building international classifications (such as ISCED; OECD, 1999) and developing measures for use in an international context (e.g. Chapter 16 in OECD, 2014b).

Yet, this is not an easy task. The first option – application of the ISCED framework – offers the potential for misclassification of educational

qualifications, as national qualifications do not always easily fit into such international classification schema (Schneider & Kogan, 2008). Moreover, this has been further complicated by the required extensive recoding of national qualifications onto the ISCED schema as well as across different versions. In contrast, developing new measures faces other difficulties, such as cultural comparability and translation issues. For instance, Rutkowski & Rutkowski (2013) showed threats to the validity and reliability of one of the socio-economic measures in PISA (the Economic, Social and Cultural Status index) due to poor cultural comparability (as well as poor model-to-data fit of subscales).

4.2.5 Measurement error in indicators of socio-economic background

The usefulness of background measures depends upon whether they manage to capture the true value of the respective measure. In addition to the difficulties mentioned above of constructing a valid measure, background measures also face further challenges if participants act as proxies for other people (e.g. students providing information about their parents). This can easily lead to misreporting which then, in turn, introduces bias into group comparisons. For instance, when students are used as proxies for their parents their answers are generally less accurate, with this also varying by factors such as age and gender (Ridolfo & Maitland, 2011). In ILSAs, socio-economic status is mainly measured by students' responses to questions about their parents' occupations and education, as well as possessions at home (only a handful of countries administer additional optional questionnaires where this information is also collected from parents). In a cross-national context, Jerrim & Micklewright (2014) showed that deviations in the answers of students and parents can lead to biased comparisons of socio-economic inequalities across countries.

4.2.6 Socio-economic measures and the construction of ILSA test scores

ILSAs only have limited time available to test students. They thus employ a rotated test design, meaning each student only gets asked a subset of questions. This results in large amounts of missing test-item data by design.

In order to estimate the achievement of sub-groups (e.g. average test scores of students from socio-economically advantaged and disadvantaged backgrounds), a complex statistical methodology (known as 'conditioning') is used. As noted in Chapters 2 and 3, this is in reality an application similar to multiple imputation. Specifically, answers to the test items actually administered are used to estimate the students' achievement distributions, with information from background questionnaires (e.g. gender, age, socio-economic status) then used to adjust the achievement distributions for population characteristics (Mislevy et al., 1992; von Davier et al., 2009). Importantly, this means that measures of socio-economic status (such as parental education) have an influence on the construction of ILSA test scores. Indeed, previous research has shown that measurement error in background variables (such as socio-economic status) can lead to bias in ILSA measures of achievement (L. Rutkowski, 2014).

4.3 Data & methods

4.3.1 Data

I used the PISA 2012 data from Germany (n = 5001 students) to illustrate potential bias that can affect socio-economic group comparisons (see Subchapter '2.2.1 Data' for more details about PISA 2012 data in general). This setting was chosen due to the uniquely rich information available for this country in this particular PISA cycle. Specifically, PISA 2012 was the last cycle to administer my measure of interest – highest parental education – in both the student and parental questionnaires. This is important, as measurement error in family background variables is likely to be one of the major sources of survey error when attempting estimate socio-economic achievement gradients using ILSAs. Yet, the number of items about parental education and the corresponding answer categories deviated between students and parents. Moreover, while questions about parental education were asked in all countries, students could actually provide more fine-grained responses than has been published in the publicly available PISA dataset (OECD, n.d.-

b).[29] This is available in a country-specific version of the PISA 2012 data, which I have obtained for Germany (Prenzel et al., 2015). Importantly, this data included much more detailed questions on parental education for parents and students in contrast to the publicly available PISA 2012 dataset. As I will illustrate, this allows us to investigate whether the coding of parental education into the ISCED framework was appropriate in the case of Germany, and the affect that this has upon estimates of socio-economic disparities in student achievement.

4.3.2 Sampling design in PISA 2012

PISA 2012's sampling design consists of two main components: the definition of the target population and the sampling procedure. The target population in PISA is defined as 15-year-olds who are enrolled in at least Grade 7. As a result, some 15-year-olds, such as home-schooled children, permanent exclusion, those who have repeated many grades and drop-outs past the school compulsory age, are not covered. While the target population in Germany is the same as the general population of 15-year-olds, this does not necessarily hold true for other countries (e.g. low school enrolment rates; Education Datalab, 2017) and could thus bias international comparisons. Furthermore, countries can decide to not cover their whole target population in the sampling. They are allowed to exclude schools and students as long as the overall exclusion rate is below 5% (OECD, 2014b, pp. 66–68). These exclusions can happen due to one of five criteria, four of which apply to all countries: 1. intellectual disability, 2. physical disability, 3. non-native students with insufficient language skills within their first year of arrival and 4. students who speak languages for which the mathematics test is not available. The fifth criterion can vary by country, at the discretion of the PISA national project manager. Among the listed exemplary reasons are remote geographical regions, language groups due to political or organizational

[29] Specifically, the parental education questions in Germany included many more options than the variable available in the publicly available PISA 2012 dataset suggests. This is due to the latter condensing Germany's national qualifications to fit into the ISCED framework.

reasons and special educational needs students, e.g. students with dyslexia (OECD, 2014b; Chapter 4).

Once the sampling criteria are defined, students get selected into PISA using a two-stage procedure. Schools are first selected and then students are randomly sampled within schools. There is also stratification in the sampling of schools in order to maximise efficiency. In Germany, schools are separated into 18 explicit strata, based upon: (a) 16 federal states; (b) all vocational schools; and (c) all special needs schools. Subsequently, within each of these explicit strata, schools are sorted by a set of implicit stratification variables (school type and school size in Germany as well as federal state for vocational and special needs schools). Schools are then sampled proportional to their size.[30] Two replacement schools are immediately drawn in case the initially sampled school refuses to participate. The replacement schools are supposed to be as similar as possible to the original school. They are identified as the schools in the same explicit strata – ranking directly above and below on the implicit stratification variables – as the originally sampled school. In PISA 2012 in Germany, within each school, 25 students were sampled randomly without replacement (OECD, 2014b; Chapter 4; Prenzel et al., 2013; Chapter 10).

4.3.3 Test design

As time is a limiting factor in ILSAs, students only get asked a fraction of all cognitive questions. This is done using a rotated test design. Specifically, students get randomly assigned to one of 13 test booklets. Each booklet contains four item clusters, which each contain multiple questions in one domain. In PISA 2012, there are 13 item clusters.[31] As mathematics is the

[30] The sampling chance of small schools (35 or less enrolled suitable students) was equal. This means that schools with 35 suitable students were as likely to be sampled as those with less. Schools with more than 35 students experienced a proportional increase in the sampling probability according to their size.

[31] Each country gets administered questions from 13 different item clusters. Countries with low expected results can opt to exchange two of the 'standard' mathematic item clusters with easier ones.

major domain,[32] seven of the 13 item clusters address mathematical topics and only three each are about science and reading. While each booklet contains at least one mathematical item cluster, the majority of the booklets only contain two of three domains – only 40% of the students answer questions in all three domains. As a result, the data of each student consists mainly of missing data by design (OECD, 2014b, pp. 30, 31).

In contrast to other PISA cycles or other ILSAs, PISA 2012 extended their rotated test design to the student questionnaire. Yet, this was done in a different way to the cognitive test. All three versions of the student questionnaire included one common part, where basic information, such as gender, language and parental education, was recorded. Fortunately, my variable of interest (parental education) was in the common part and therefore, administered to all students. As a result, all missing data in parental education was due to either students not answering the question or not returning the questionnaire.[33]

As I am interested in student achievement and parental education in this chapter, I am using students' answers in the three core cognitive domains, the student questionnaire, and the parental questionnaire. While the core cognitive domains and student background questionnaire are administered by default, the parental questionnaire is an additional domain which only a few countries, including Germany, choose to administer.

4.3.4 Measure of interest

In this chapter, I am examining potential bias that can affect (socio-economic) group comparisons in ILSAs. While socio-economic status has multiple facets, for the sake of comprehensibility and feasibility I focus on one aspect, parental education, for several reasons. First, parental education is a measure of socio-economic status which is commonly used in secondary analyses of

[32] On a rotation basis, in each cycle one domain receives special attention. As a result, more cognitive question are asked in that year and additional information about it is collected in the student background questionnaire.

[33] Each questionnaire booklet also included rotated parts, which covered other topics such as their experience with mathematics and their school. If one is interested in those variables for group comparisons, appropriate handling and implications due to this limited information by design need to be considered.

ILSA. In practice, it is used as a main variable as well as a covariate and is thought to be related to academic achievement (e.g. Hansen & Gustafsson, 2016; Jerrim & Macmillan, 2015; Martins & Veiga, 2010). Second, the aim is to highlight the issue of total survey error in ILSA group comparisons and I am not specifically interested in the theoretical implications of the results with respect to transgenerational achievement. Finally, the data in PISA 2012 enables us to comprehensively investigate bias in parental education gaps in academic achievement.

The publicly available PISA 2012 data contains students' answers about their parents' education in two subtopics, schooling and professional education, each for mother and father. Using responses to these questions, the OECD derives the ISCED 1997 level for mother and father by recoding the answers for national qualifications onto the ISCED scale. The highest parental education overall is then computed by combining categories and the responses for mother and father.[34] In contrast to the student questionnaire, the parent questionnaire only mainly enquires about the parents' professional education and not about their schooling, i.e. no information on education below ISCED level 3A – see Table 4.1. Furthermore, the items are not the same as in the student questionnaire – at least in the official published version. As a result, highest parental education based on the student and parent questionnaire from the publicly available data cannot be compared directly (compare column Student Int. and Parent Int. in Table 4.1).

The German data is available upon request (Prenzel et al., 2015) and has more fine-grained questions about parental education than the international version (the ISCED levels in Table 4.1 are already combining categories in some cases). While the questions are still structured into schooling and professional education, each topic has more options than the international version. The German school and professional education system is complex with multiple tracks and an extensive apprenticeship system which cannot be briefly explained. For example, after primary school (usually Year 4), students are usually transferred into one of three tracks (lowest track: Hauptschule – 9

[34] This information is available as a composite measure in the dataset. The procedure is outlined for understanding.

years, middle track: Realschule – 10 years, highest track: Gymnasium – 12/13 years), which have different orientations and lead to different qualifications and ISCED levels. For a comprehensive summary of the German system and ISCED level classifications, see Schneider (2008).

Table 4.1 Available information on parental ISCED levels in the different background questionnaires and versions

<table>
<tr><th colspan="3">Student</th><th></th><th colspan="3">Parent</th></tr>
<tr><th>Ger.</th><th>Int.</th><th>CM</th><th></th><th>Ger.</th><th>Int.</th><th>CM</th></tr>
<tr><td colspan="2">Did not complete ISCED 1</td><td></td><td></td><td>Did not complete ISCED 1</td><td rowspan="4">ISCED 3B, C or below</td><td></td></tr>
<tr><td colspan="2">ISCED 1</td><td></td><td>ISCED 1</td><td></td></tr>
<tr><td colspan="2">ISCED 2</td><td></td><td>ISCED 2</td><td></td></tr>
<tr><td colspan="2">ISCED 3B, C</td><td></td><td>ISCED 3B, C</td><td></td></tr>
<tr><td colspan="2">ISCED 3A</td><td>ISCED 3A</td><td colspan="3">ISCED 3A</td></tr>
<tr><td colspan="2">ISCED 4</td><td>ISCED 4</td><td colspan="3">ISCED 4</td></tr>
<tr><td colspan="2">ISCED 5B</td><td></td><td colspan="3">ISCED 5B</td></tr>
<tr><td colspan="2">ISCED 5A</td><td rowspan="2">ISCED 5A
ISCED 6</td><td>ISCED 5A</td><td rowspan="2">ISCED 5A, 6</td><td></td></tr>
<tr><td colspan="2">ISCED 6</td><td>ISCED 6</td><td></td></tr>
<tr><td>Other</td><td>-</td><td>-</td><td>Other</td><td>-</td><td>-</td></tr>
</table>

Note: Ger. = German Items. Int.= International Items. CM = (International) Composite Measure. ISCED levels correspond to: 1 – Primary education, 2 – Lower secondary education, 3 – Upper secondary education, 4 – Post-secondary non-tertiary education, 5 – First stage of tertiary education, 6 – Second stage of tertiary education. The letter after the number depends on to which further level it grants you access. All international items were valid for Germany, but in some countries certain ISCED levels do not occur/were not asked for, e.g. ISCED level 5B in Poland.

In contrast to the international version, the German data has the major advantage that the same questions and response options were posed to students and their parents, i.e. I can directly compare their responses. Yet, the pre-computed highest parental education variable (in ISCED levels) in the German data does not make use of the more-detailed questions in the German version. Rather, it classifies highest parental education into the fewer and broader categories from the international version. The mapping of the German responses to ISCED levels (in PISA) is outlined in the official PISA scaling manual for Germany (Mang et al., 2018, pp. 173–176). The exact questions and details of highest parental education can be found in Appendix C.1.

4.3.5 Method of plausible value computation

As described previously, not all students answer questions in all three core domains. Yet, PISA assigns every student a score in mathematics, reading and science. In order to do this, complex multi-step statistical procedures are needed. Thereby, background variables (such as gender and highest parental education) are used in combination with the cognitive items to derive student achievement distributions (OECD, 2014b). Thereby, student achievement distribution is estimated through an IRT model using the students' cognitive responses. Subsequently, prepared background variables are used in a latent regression, the so-called 'conditioning model' to adjust the distributions for population characteristics. Subsequently, plausible values are randomly drawn from the resulting student achievement distribution.

The aim of this chapter is to illustrate the potential sources of bias in group comparisons. In Subchapter '4.4.6 Plausible value computation', I analyse the impact that including highest parental education in the plausible value computation can have upon estimates of the parental education achievement gap. As a result, I use a simplified and altered version in comparison to PISA to analyse the impact that highest parental education can have through the plausible value computation process. I conduct three different versions of the conditioning model – 0. No variables in the latent regression, 1. Only highest parental education based on the students' responses (no other variables) in the latent regression and 2. Only highest parental education based on the students' responses (no other variables) – and draw plausible values afterwards. The resulting different sets of plausible values are then used to investigate the impact. For further details of plausible value computations, see Subchapters 2.2 and 3.2.2. For further information on the practical implementation of this chapter, see Appendix C.2

4.4 Analysis

All analyses in this chapter are conducted within the statistical programming language 'R' (R Core Team, 2019). In order to account for the design in PISA and compute valid scores, the 'intsvy' package (Caro & Biecek, 2017) is used

together with the Balanced Repeated Replication (BRR) weights and the final student weight.

4.4.1 Coverage of the (target) population

The target population in PISA 2012 is defined as all 15-year-olds which are enrolled at least in Grade 7. The sampling framework allows countries to not cover parts of their target population due to a few reasons; see Subchapter '4.3.2 Sampling design in PISA 2012' for more details. Key information and numbers about the conducted sampling are published by the OECD with an international focus (OECD, 2014b; Chapter 4, Chapter 11) and the national centres with a focus on the national situation (Prenzel et al., 2013; Chapter 1.3.1).

The German target population equals the number of all 15-year-olds in Germany (N = 798,136). Overall, the number of (weighted) excluded students due to school exclusion was 10,914 or 1.4% of the target population. In the sampled schools, eight students were excluded and 56 were withdrawn or deemed ineligible[35], which corresponds to a weighted estimate of 1,302[36] and 7,805 in the whole target population or 0.2% and 1.0% respectively. The number of not covered schools and students accumulates to a non-coverage rate of 2.6%. Thinking back to the reasons for exclusion (e.g. language, special educational needs), it is likely that exclusion is not random but related to achievement. While a (direct) relationship to highest parental education cannot be definitively proved (due to a lack of data), an association between parental educational background and student achievement is often found (e.g. Gamboa & Waltenberg, 2012; Ludeke et al., 2021).

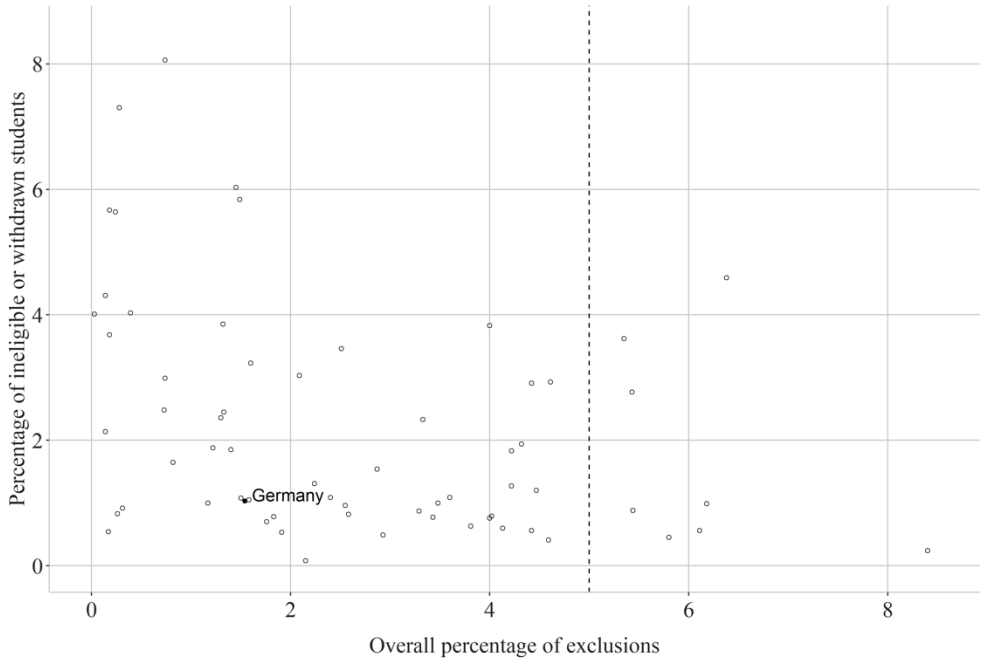
Nevertheless, Figure 4.1 shows that this non-coverage rate is small in comparison to most other countries in PISA 2012. It clearly highlights that in most countries a larger proportion is either excluded or withdrawn/deemed ineligible than in Germany. Eight countries (Canada, Denmark, Estonia,

[35] Students were considered ineligible if they were age-eligible first, but subsequently not meeting other definitions of the target population, e.g. left school after sampling but before the assessment.

[36] This number varies slightly between the German (1,357) and the international (1,302) report but both documents state the same percentage of the total population.

Luxembourg, Norway, Sweden, the United Kingdom, and the United States) even exceed the official threshold for the overall exclusion rate (yet have still managed to pass to OECD's data adjudication process). While non-coverage rates should never be ignored and can always have an impact on results, especially if systematic, other countries are at substantially higher risk of bias than Germany due to significantly higher non-coverage rates.

Figure 4.1 International comparison of the overall exclusion rate and the 'deemed ineligible or withdrawn' rate in Germany



Note: Dashed line refers to the rate threshold of acceptable exclusions according to the official OECD documentation. Solid point represents Germany, while the hollow points represent the other countries participating in PISA 2012.

4.4.2 Survey non-response

When schools and students are sampled, there is no guarantee that all participate. While public schools in Germany were obligated to participate in PISA 2012, private schools could decide whether they wanted to participate or not. According to Prenzel (2013), five private schools declined to

participate in PISA 2012. Two of these were replaced. Overall, 230[37] of 233[38] schools participated (after replacements included) resulting in a school response rate of 98.7%. I cannot test for any relationship between the three private schools that declined to participate and highest parental education, as I do not have any information about those schools, but Lohmann et al. (2009) showed that parental education is a central factor whether a child goes to a private school in Germany.

Student non-response (in the cognitive domains) can stem from different reasons such as sampled students were unavailable at test day or difficulties during administration. The realised student sample for Germany in PISA 2012 encompassed 5001 students.[39] Using the number of 5355 sampled students from the technical report (OECD, 2014b, p. 185), this leads to a student response rate of 93.4%.[40]

As with the coverage rates, Germany does fairly well in international comparison of these response rates. School response rates (unweighted after replacement) vary internationally between 77.8% in the United States and 100% in multiple countries in PISA 2012. Similar variability between countries can be found for student participation: Canada had the lowest student participation rate with 80.8% – barely meeting the required threshold of 80%. In contrast, in Vietnam 99.9% of the sampled students participated. In general, it is especially troublesome if both rates are low, as in the Netherlands in 2012 (student response rate: 85.0%, school response rate: 88.9%). While Germany has comparatively high participation rates, it does

[37] The technical report states 228 schools after replacement for Germany (OECD, 2014b, p. 183). I decided to use the number of the German report, as it coincides with the number of schools in the data.

[38] The sampling frame encompassed 247 schools, but 11 schools were dropped completely because they did not have any 15-year-olds, one school did not exist anymore and one school was excluded due to language feasibility and should therefore (hopefully) be included in the coverage rate in the previous subchapter (Prenzel et al., 2013, pp. 319, 320).

[39] The number of participating students in Germany is stated either as 4990 or 5001 in the official technical report depending on the page (OECD, 2014b, pp. 178, 185). I decided to use 5001 students, as it coincides with the number of students in the German report and the dataset.

[40] The German report states a student response rate of 93.2% but without any statement of the number of sampled students (Prenzel et al., 2013, p. 320).

not necessarily mean that little or no bias is introduced, though the risk is lower than elsewhere.

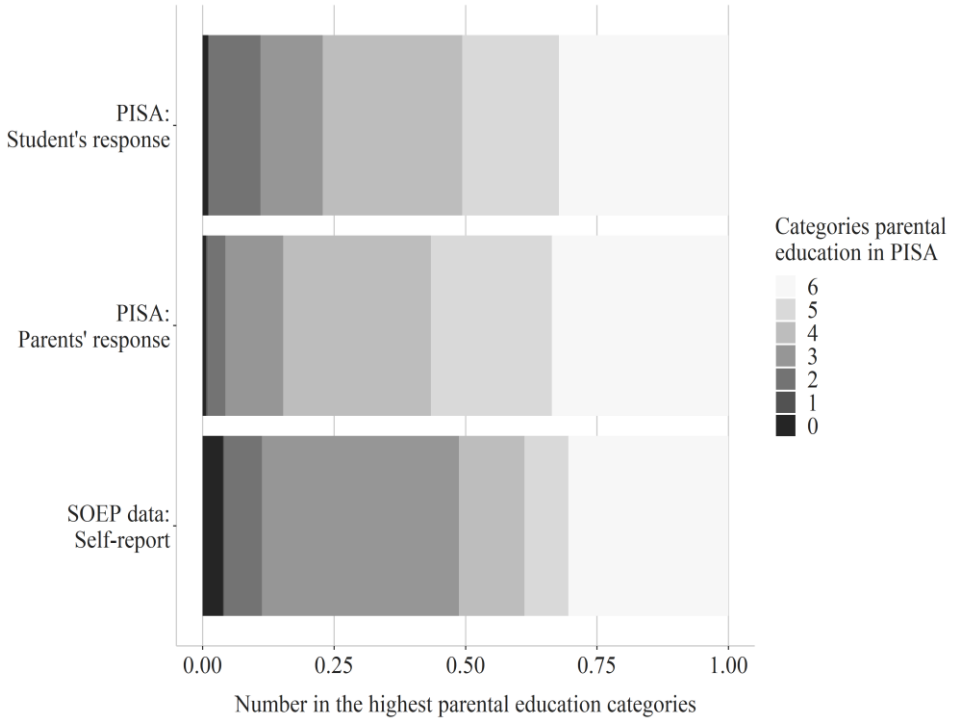
PISA also aims to collect additional information via student questionnaires. While I do not have any information on students that declined to participate in the PISA test, it is possible that students who took the test did not complete the background questionnaire. The 16 German states enjoy educational sovereignty and handled the administration of the student questionnaire differently: five states made the student questionnaire obligatory, 10 states opted for voluntary return and required parental approval and the final state decided that parts of the questionnaire were compulsory. As a result, response rates in Germany differed between states (Prenzel et al., 2013, p. 33). Additionally, different states in Germany are also associated with different demographic structures (Statistische Ämter des Bundes und der Länder, 2011). Overall, the data contained 684 of 5001 cases with missing data only in the questionnaire items. This results in a student questionnaire return rate of 86.3%.[41] When combining the general student response rate with the student questionnaire return rate, 80.6% of the sampled students answered and returned both key parts of PISA. Germany chose to administer further additional questionnaires, such as the parental background questionnaire. The return of the parental background questionnaire was voluntary in all states. Overall, 2885 questionnaires were returned – a parental background questionnaire return rate of 57.7%.

Figure 4.2 compares the distribution of highest parental education in PISA 2012[42] to the distribution of highest household education in the German socio-economic panel (SOEP) data in 2012 (Socio-Economic Panel, 2019). While the SOEP sample is different to PISA, as it covers the whole population, steps were taken to align the samples as closely as possible[43]. The

[41] The German report states a return rate of 82% but without any further breakdown/numbers and the official and publicly downloadable data contains information for more cases.
[42] Some responses are also missing due to item non-response, but missing data is driven in large parts by survey non-response, especially for the parents' responses.
[43] Highest education was calculated for the adults, born between 1952 and 1982, of each household where an adult had at least lived in a household with a child at one point. This was done in order to get as close as possible to the sample of parents who could have had a 15-year-old in 2012. The SOEP variable 'pgisced97' was slightly recoded to match highest parental education in PISA.

figure shows large differences between the PISA distributions and the one based on SOEP data. While the highest category stays roughly the same across all rows, in all other categories there are substantial differences. This is most prominent for Category 3 (ISCED 3B, C), which is mainly corresponds to apprenticeship in Germany. In PISA, only 12% of the students and 11% of the parents indicated that this was their highest education. In contrast, apprenticeship amounts to 37% in the SOEP sample. While the information based on the SOEP data clearly deviates from the PISA distribution, it is not necessarily unrealistic. The German statistical yearbook in 2012 (Statistisches Bundesamt, 2012; Chapter 3) states that an apprenticeship is the highest professional education for roughly half of the population. While that amount will decline slightly when looking at highest education in general (there are cases where the educational level of schooling is higher than an apprenticeship), the amount of apprenticeship as highest education will still exceed the 37% in the SOEP sample. Overall, while the SOEP data is not a representative sample of parents of 15-year-old students in Germany, it gives sufficient indication that the response and return rates in PISA are biased in terms of socio-economic background.

Figure 4.2 Distribution of highest parental education in Germany based on (i) student's answers in the PISA dataset, (ii) parents' answers in the PISA dataset and (iii) household data in the German SOEP dataset



Note: Categories: 0 = Below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. The student sample is based on 3936 students from the German PISA 2012 dataset (with the parental education variable coded as described in '4.3 Constructing and coding a valid socio-economic group measure'). The parent sample is based on 2832 parental background questionnaires from the German PISA 2012 dataset (with the parental education variable coded as described in '4.3 Constructing and coding a valid socio-economic group measure'). The SOEP data sample is displaying highest education in 19629 households, which were selected according to the following criteria: (a) person was not interviewed as child or youth of the household, (b) person born between 1952 and 1982, (c) one household member lived in a household with at least one child at one time point. Minor recoding of the SOEP variable 'pgisced97' was necessary in order to align it with the highest parental education in PISA. The planned assignment of parental highest education is used in this figure instead of the initially observed one (see Subchapter 4.4.3 for more information).

4.4.3 Constructing and coding a valid socio-economic group measure

In order to validly measure parental education in an international large-scale assessment, education needs to be on an internationally comparable scale. In PISA 2012, this was the ISCED 1997 scale. In the most straightforward case, students and parents would classify their education correctly as ISCED

categories. Yet, this is not feasible, as they are mostly only aware of their country-specific qualification and not the international 'equivalent' (ISCED classification). As a consequence, they are asked about the country-specific – German in this case – qualifications. These questions then need to be 'translated' or more precisely recoded to the ISCED scale in order to facilitate international comparisons. The precise procedure for this in Germany is outlined in Mang et al. (2018, pp. 173–178).

Errors in processing and coding in such situations are known as procedural errors in the total survey error framework. Indeed, when checking the (re)coding of the German questions to the international items, it became clear that a systematic error in the (re)coding occurred, as highlighted in Table 4.2. Importantly, these errors are present in the PISA 2012 public use PISA data files for Germany, downloadable from the OECD website.

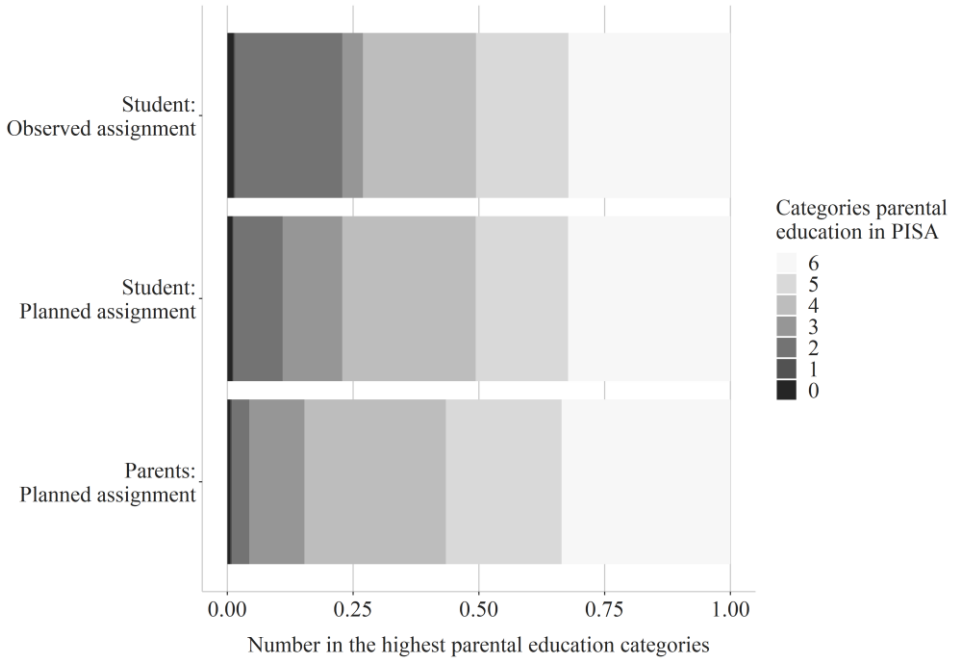
Table 4.2 Procedural error in highest parental education in the case of Germany: Comparison of the planned and observed assignment

Variables in the German dataset	**ISCED level**	**Planned assignment** (Mang et al., 2018)	**Observed assign-ment***
DEU_ST1^{X}N01ϵ {8,9}	Below ISCED 1	0	0
DEU_ST1^{X}N01 = 6	ISCED 1	1	1
DEU_ST1^{X}N01 ϵ {3, 4, 5}	ISCED 2	2	2
DEU_ST1^{Y}A06 = 1	ISCED 3B, C	3	**Ignored.**
DEU_ST1^{X}N01 = 1	ISCED 3A, 4	4	4
DEU_ST1^{Y}N05 = 1	ISCED 3A, 4	4	**3**
DEU_ST1^{Y}N04 = 1	ISCED 5B	5	5
DEU_ST1^{Y}N03 = 1	ISCED 5A, 6	6	6
DEU_ST1^{Y}N02 = 1	ISCED 5A, 6	6	6
DEU_ST1^{Y}N01 = 1	ISCED 5A, 6	6	6

Note: ^{X} corresponds to 3 for the mother and 7 for the father. ^{Y} corresponds to 14 for the mother and 18 for the father. The highest parental education is the highest score of mother and father. *These are the conclusions derived from comparing the variables in the German dataset to the recoded items in the international dataset. As a result, I cannot be 100% sure if this is what happened. According to the data, there were also a few minor exceptions where the values were set to missing due to plausibility checks, e.g. if no question about schooling was answered and no 'lower' professional education was ticked, but PhD was ticked as highest education.

As a result, while the two highest categories remain unaffected, the other categories experience changes as highlighted in Figure 4.3. The bar chart in the first row shows the proportions for highest parental education based on student answers as they can be found in the international PISA dataset (i.e. the observed assignment). In the row below, the corrected scale according to the official documentation (i.e. the planned assignment) can be seen. The comparison of these highlights how the proportion of parents in Category 3 increased substantially, while the proportion in Category 2 was halved, if highest parental education is properly coded. Naturally, this will impact group comparisons, as the proportion of students within each category has changed substantially. In the last row, the bar chart of the highest parental education based on answers from a parent is displayed. There is no counterpart in the international dataset as it does not exist in the same detail, because the international parental questionnaire did not include the necessary questions. The more detailed German data allows us to compute a composite measure which can be compared to the measure for highest parental education based on the student answers.

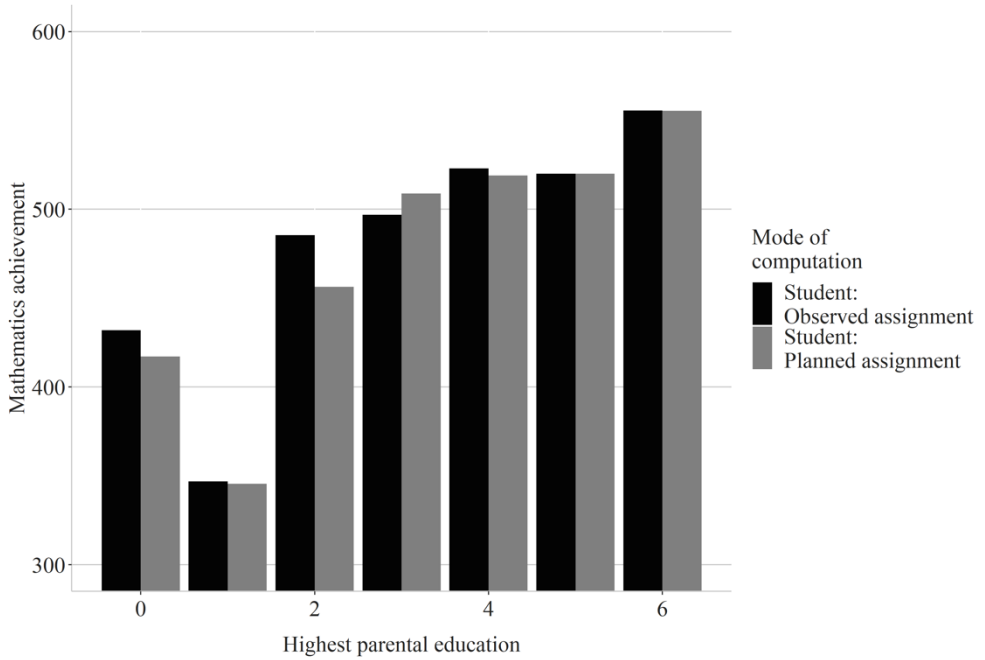
Figure 4.3 Distribution of highest parental education in Germany for (i) observed students' answers in the international dataset (with procedural error), (ii) students' answers according to the planned assignment and (iii) parents' answers using the planned assignment



Note: Categories: 0 = Below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. Sample size students = 3936. Sample size parents = 2832. The first row displays the observed assignment of highest parental education (student response) in the German PISA 2012 data. The second row displays the planned assignment according to the German scale manual (Mang et al., 2018) based on student responses. The third row displays the planned assignment according to the German scale manual (Mang et al., 2018) based on parental responses**.**

The procedural error also impacts the average mathematics PISA scores for each parental education group, as shown in Figure 4.4, with all affected categories experiencing change. With the exception of Category 3 (ISCED level 3B, C), mathematics achievement dropped. Thereby, Category 2 (ISCED level 2) had the largest decline with 29 points, while the average achievement rose by 12 points in Category 2.

Figure 4.4 Comparison of average mathematics achievement of the highest parental education group based on the observed scale in PISA and based on the planned (and corrected) assignment



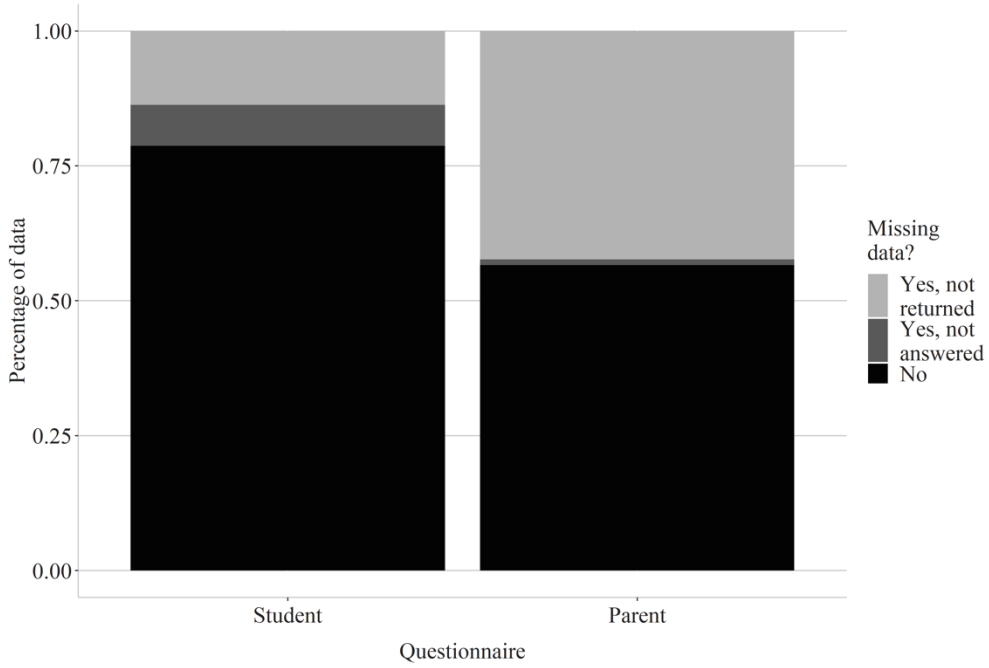
Note: Categories: 0 = Below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. Group means were computed using plausible values and accounting for sample weights.

In the international report in PISA 2012, one important predictor of disparity between socio-economic groups is the difference in achievement of students with highly skilled parents in comparison to students with low-educated parents (OECD, 2013, pp. 38, 40, 189). This is computed through the mean difference between students with at least one parent who completed ISCED Level 5 or 6 (Categories 5 & 6) and students where the highest parental education is ISCED Level 2 (Categories 0-2). The mean difference is 62 points in Germany when using the observed scale from the official published PISA dataset, whereas it is 92 points when the errors are corrected. This leads to a substantial underestimation of the gap between students with high-educated parents and low-educated parents by 30 PISA points (or, equivalently, the gap has been underestimated by almost 50%). When using the flawed scale, Germany ranks 24th in an international comparison of parental education achievement gaps (OECD average = 77 points), but it drops substantially to 56th place based on the corrected scale.

4.4.4 Item non-response

For roughly 20% of the students, no data is available about highest parental education. Overall, Germany has the highest proportion of missing data (on student reported parental education levels) by far. The next highest amount of missing data occurs in the Netherlands with 8.6% and the OECD average is substantially lower (3.8%). But, on the upside, Germany administered the parental background questionnaire, so additional information is available in comparison to most other countries, which can also be used to investigate missing student data further. The missing data in Germany can occur through two pathways: (i) the student and parent(s) did not return the questionnaire – survey non-response covered in Section 4.4.2 – or (ii) items relating to parental education were not answered validly or skipped – item non-response. This is illustrated in Figure 4.5, highlighting the different pattern between students and their parent(s). In general, survey non-response is the prevailing reason for missing data. It becomes obvious that far fewer parental background questionnaires were returned, but most (98%) of those that did provided a response to the question about parental education. For the students, I have substantially more information available overall (for 79% of observations highest parental education data is available). The missing data (the remaining 21% for students) had a higher proportion of missingness due to item non-response than for the parents – but still about a third of missingness was caused by item non-response (and two-thirds by survey non-response).

Figure 4.5 Percentage of different types of missing data for highest parental education in the student and parent background questionnaire



Note: Student sample is 5001 with 381 item non-response and 684 survey non-response. The parental sample is 5001 with 53 item non-response and 2116 survey non-response.

Table 4.3 shows that mathematics achievement is higher on average in the cases where students or parents answered questions about highest parental education. Achievement was at least 36 points (or more than a third of the PISA standard deviation) lower when no information was available. Furthermore, in both cases item non-response, i.e. actively choosing not to answer the questions relating to parental education, was associated with even lower scores than survey non-response, i.e. choosing to not return the questionnaire at all. By far the biggest difference is between parents with valid answers and parents with item non-response. While the mean achievement is based on a small sample size (n=53)[44], it is still indicative of the direction and magnitude of potential bias and fits the overall picture.

[44] The sample size is small, as most parents answered all questions if they chose to return the questionnaire.

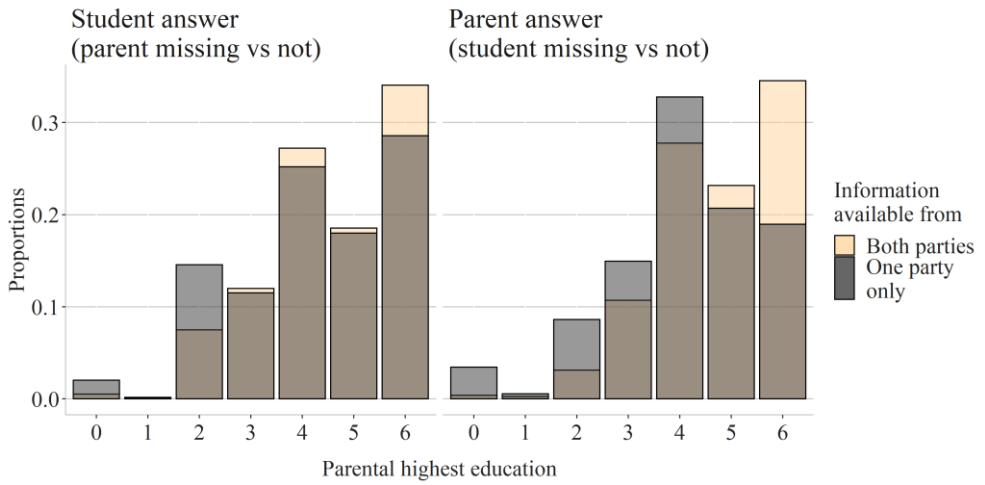
Table 4.3 Mathematics achievement conditional upon whether questions about parental education were answered or survey or item non-response occurred

	Average mathematics achievement when the	
	student response is…	**parent response is…**
… available	522	532
… missing (survey non-response)	486	490
… missing (item non-response)	473	448

Note: Group means were computed using plausible values and accounting for sample weights. Sample sizes from top to bottom, then left to right: 3936, 684, 381, 2832, 2116, 53.

Another way of looking at potential bias is comparing the distribution of highest parental education based on students' responses where I have information from both students and parents to those where I only have the students' responses – and vice versa. Overall, I have information from both parties for half of the cases. For roughly another quarter (28%), I have information from the student but not the parents. The opposite direction (parents answered questionnaire and items, but students did not) was only present in 3% of the cases. Figure 4.6 displays the differences in the distribution of highest parental education between these groups (response of the common party is used, e.g. student answer for response from student only versus response from student and parent available). It becomes obvious that the distributions differ depending on whether one or both parties have information available, with most categories experiencing noticeable deviations. In general, parental highest education was lower if either the parent or student did not provide an answer. The proportion of students with high-educated parents was higher if both parties answered, whereas the proportion of students with low-educated rose if only one party answered. This is a strong indicator that missingness does not occur (completely) at random. Estimates of socio-economic inequalities based upon this variable are hence likely to be subject to some bias, though the exact magnitude of this bias is hard to determine with the data available.

Figure 4.6 Distribution of highest parental education based on whether information is available from both parties or just one



Note: Categories: 0 = Below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. The sample size of students with information about parental education from both parties is 2658, with information from students only is 1278 and with information from parents only is 174. The planned assignment of parental highest education is used in this figure instead of the initially observed one.

4.4.5 Measurement error: Agreement of parents and students

Germany was one of only 10 countries that administered the parental background questionnaire in PISA 2012. As they asked students and parents the same questions about parental education, I can evaluate the level of agreement between those two. Table 4.4, a cross-tabulation between parent and student responses, shows that students do indeed report their parents' education level with error (assuming that parents know and reveal their 'true' level of education). Out of the 2658 cases where I have information from both parties, only 1376 – slightly more than half of them – selected the same category. In total, 540 students assumed that their parent(s) had a higher level of education than they stated, while the other 742 thought the opposite. In general, there is no clear evidence in Table 4.4 that students provide an answer 'close' to (i.e. within one category) of their parents' answer. Overall, the polychoric correlation is 0.656, with the Kappa statistic (a measure of inter-rater reliability that takes into account chance agreement) standing at 0.36. While these values are generally considered fair, they mainly refer to

agreement over single items in a test. Parental education is used as a grouping variable and to derive measures of inequality. These measures can change dramatically depending on the respondent. Therefore, to get robust and consistent results it would be desirable that agreement between students and parents is higher and thus realistically depicts their families' socio-economic status.

Table 4.4 Crosstab of the students' and parents' responses for highest parental education

		Student response								
		0	1	2	3	4	5	6	Total	NA
Parent response	0	7	1	1	0	2	0	0	11	6
	1	1	2	2	0	0	1	1	7	1
	2	3	0	31	20	14	8	7	83	15
	3	1	1	42	83	104	40	14	285	26
	4	1	1	69	148	313	131	75	738	57
	5	0	0	37	59	150	251	119	616	36
	6	1	0	17	9	140	62	689	918	33
	Total	14	5	199	319	723	493	905	2658	
	NA	26	2	186	147	322	230	365		894

Note: Categories: 0 = Below ISCED 1; 1 = ISCED 1; 2 = ISCED 2; 3 = ISCED 3C, 3B; 4 = ISCED 3A, 4; 5 = ISCED 5B; 6 = ISCED 5A, 6. NA stands for missing data. The planned assignment of parental highest education is used in this table instead of the initially observed one.

Returning to the previous illustration of bias using the mathematics gap between students with high- and low-educated parents, there are changes depending on whose response for highest education is used. After restricting the sample to those with valid information on highest parental education from both students and parents, the gap in mathematics achievement is 89 points using students' responses, but 116 points when using parental responses. This is an increase of more than 25 points (approximately a quarter of a standard deviation). Assuming that parents know their own education better than their children, this shows that the gap is severely underestimated when using students' reports. Yet, Germany is one of the few countries where information from both parents and students is available at all, and where in turn the difference can be investigated. Even for the countries which administered the parental questionnaire, comparable ISCED measures based on students' and

parental response are not available in the official published dataset. Similar bias may occur in other participating countries but cannot be detected.

4.4.6 Plausible value computation

In addition to other sources, plausible value computation can also lead to bias. The conditioning model (see '4.2.6 Socio-economic measures and the construction of ILSA test scores' and '4.3.5 Method of plausible value computation' for details) is used to counteract possible attenuation in group comparisons. As a result, when conditioning is not conducted or based on incorrect information, bias is introduced.

I illustrate this using an oversimplified exemplary plausible value computation and subsequent analysis: first, computing three sets of plausible values (including or excluding different information on highest parental education) and second, analysing the gap between students with high- and low-educated parents. Thereby, I want to show the difference in gaps that stems just from using the students' responses to the parental education questions versus using the parental responses in the conditioning model. Again, I reduced the sample to students with information about highest parental education from both parties. I then compute the three different sets of plausible values based on three different conditioning model specifications:

0. No conditioning
1. Only highest parental education based on the students' responses (no other variables)
2. Only highest parental education based on the parental responses (no other variables)

After drawing the plausible values, the gap in mathematics achievement between students with high- and low-educated parents is computed and stated in standard deviations,[45] which are shown in Table 4.5. The first column

[45] The gaps are reported in effect sizes and not PISA points because the plausible values are not automatically computed on the PISA scale and correctly transforming them on the scale would be disproportionately time-consuming and complex. The standard deviation of mathematics in PISA 2012 is 96 points for Germany.

shows the gap when the grouping (high-/low-educated parents) is based on students' responses, while the second grouping uses parental responses. The rows correspond to the different conditioning model specifications. As expected, the effect sizes are smallest when no conditioning is applied and rise as soon as any measure of highest parental education is included in the model. Yet there are differences between the results from conditioning model (1) and (2) which are solely attributed to the different variables in the latent regression. For the gap based on students' responses, the impact is 0.02 standard deviations (or around 2 points in PISA) and for the gap based on parental responses it is 0.11 standard deviations (or around 11 points in PISA). Overall, this shows that it matters both (a) whether the grouping variable (parental education in this instance) is included in the conditioning model and (b) whether there is potential measurement error within this variable.

Table 4.5 Difference in standard deviation between students with high- and low-educated parents based on different conditioning models and grouping variables

Conditioning model	**Grouping variable based on**	
	Students' response	**Parents' response**
0. No conditioning	0.72	1.10
1. Students' response	0.75	1.17
2. Parents' response	0.77	1.28

Note: In the first row, no conditioning was used, i.e. only IRT. In the other two rows, the conditioning model only included a single variable each: Highest parental education (corrected scale) based on the students' response in the second row and based on the parents' response in the third row. Only students who had valid information from both parties were included in the analysis. The effect size is the mean difference between students whose parents had at least ISCED 5 as highest education and those with parents with ISCED 2 as highest education and divided by the pooled standard deviation. The classification between high- and low-educated parents was either done based on the student's response or parent's response. The planned assignment of parental highest education is used in this table instead of the initially observed one.

4.5 Conclusions & recommendations

The influence of ILSAs, such as PISA, has increased over time. They have long become an important part in national and international education debates

and as a result, a tool in education policy decision-making. Therefore, it is of great importance that the data underlying ILSAs has a solid foundation. Yet, this is not always guaranteed. The data and methodology can be biased at different points in the process, from sampling through to the derivation of the plausible values. This, in turn, may bring into question the robustness of the results.

While there are many studies focusing on one single (potentially) problematic aspect of PISA, few have taken a broader, more comprehensive perspective. As a result, investigations are more in-depth, but fail to show the total picture of things to consider and set into reference. The aim of this chapter has been to start to fill this gap, thus fostering a deeper understanding about various psychometric and statistical properties of PISA. This has been done within the total survey error framework, identifying multiple aspects that can potentially impact upon country measures, via a case study of socio-economic status differences (measured via highest level of parental education) in Germany.

Specifically, six different statistical and psychometric properties of PISA have been considered. First, the target population and its sampling have been inspected in terms of non-coverage rates. Second, school response and student attendance as well as questionnaire return rates are analysed to judge quality of the sample selected. Third, the construction and coding of the ISCED measure has been checked for underlying (procedural) errors. Fourth, I have investigated whether there is evidence of selective item non-response to questions students are asked about their parents' education. Fifth, the agreement between students and parents about parental education levels has been investigated to consider whether measurement error may affect the results. Sixth, I have also investigated whether (potentially) biased socio-economic background measures impact PISA test scores solely through the computation of plausible values.

My analyses show that some – although not all – of the above have a substantial impact. For instance, comparing the (more detailed) German national data on parental education to the (less detailed) OECD categorisation

illustrated how a procedural error occurred when information was recoded into the ISCED scale. As a result, many parents were falsely classified as low-educated even though they completed an apprenticeship, widely considered to be a good, solid level of education in Germany. This affects all analyses using parental education, e.g. the gap between students with high- and low-educated parents rose from 62 to 92 PISA points due to this coding error alone. Another aspect is survey and item non-response rates. Just above 80% of the sampled students completed the test and the questionnaire, with some also missing data to single items. Analyses showed that student achievement is typically lower when ISCED data is missing. The same holds true for parental education of students with information from only one party in comparison to students with information from students and their parents. Where the information was available, students' and their parents' responses were used to compare their agreement on parental education. Just about half of them agreed and no clear pattern was found to explain the differences. Yet, if used as grouping, it is vital that students act as good proxy respondents for their parents. Yet this is not the case, meaning that, in Germany, the gap between students with high- and low-educated students differs by 25 points (0.25 standard deviations) depending upon whether student or parental reports are used. On the other hand, although the computation of the PISA plausible values can lead to some bias in parental education group comparisons, the impact of these were relatively small. Nevertheless, my overall interpretation of results is that comparisons of achievement differences by parental education level in Germany using PISA do not appear to be particularly robust.

ILSAs, such as PISA, are large and complex studies which administer hundreds of test and background questions to students. I acknowledge that it is not feasible to examine every background variable in each country as closely as I did in this case study. Nevertheless, this study poses important questions about the reliability of background variables in PISA and whether they should be used as 'blindly' as they currently are.

This leads me to two key recommendations. First, better communication of potential issues and troublesome aspects is needed. The limitations about the

background data should be more thoroughly investigated and the caveats articulated. This should ideally be accompanied by a brief guide about different psychometric and statistical aspects as well as ways to check for potential pitfalls. Second, many researchers often come across issues with the background variables they use. As a result, they either decide to use another variable or briefly note it somewhere in their paper. It would be of great benefit if such information could be easily shared with other data users. One idea would be a log or database on the PISA data homepage, where concerns can be raised unbureaucratically. Whether these claims are maintained or verified by the OECD or not, it would be valuable information to other researchers, especially for countries or variables that they are unfamiliar with.

While this chapter aims to show different sources of bias and is set in the framework of total survey error, I do not claim to look at all potential issues with the data. This study serves as a review and attempts to shed light on some infrequently discussed psychometric and statistical properties that can potentially influence secondary analyses and results. A single-country case study has been presented to facilitate an in-depth investigation, with different aspects likely to change in other settings. Nevertheless, this chapter highlights the importance of conducting such detailed investigations and how they can be systematically approached.

5 Conclusions

5.1 Key findings

The Programme for International Student Assessment (PISA) is a triennial international large-scale assessment (ILSA) that aims to comparably assess student achievement of 15-year-olds all around the world and over time in mathematics, science and reading. It has long become an important tool of soft governance and influences the discussions about education, policy decisions and daily school life. As a consequence, the PISA scores and results as well as the underlying methodology must be valid and sound. Yet, criticism surfaces time and again. While some raise ethical concerns and doubts about the interpretation of the results in general, several methodological aspects have also been questioned. This thesis adds to this literature, providing new insights into different statistical and psychometric properties of PISA.

Chapter 2 'Conditioning: How background variables can influence PISA scores' closely scrutinises the process of achievement score estimation in PISA. This is a multi-step procedure, called 'conditioning', and is a combination of an item response theory model and latent regression. I focus on investigating the impact that background variables can have on PISA scores via the conditioning model. As a starting point, I try to replicate the PISA 2012 plausible values as closely as possible using the official documentation and the published data. Afterwards, I systematically vary the background variables included in the conditioning model to gauge the impact they have upon PISA scores. My key findings are:

- The official technical documentation is not detailed enough to allow exact replication of the scores. This lack of open science does not help facilitate independent attempts to scrutinise the conditioning model and its properties.
- Close replication of PISA country average scores was successful in the major domain, but only to a lesser degree in the minor domains, especially reading.
- The exact specification of the conditioning model matters:

    - Country averages are robust for mathematics and science, but background variables have more impact on reading.
    - The 90^{th}–10^{th} percentile gap, a common measure of inequality, is substantially impacted by the conditioning model specification in all three domains.
    - Gender gaps are robust, with very few exceptions, to the exact specification of the conditioning model – as long as gender is included.
    - The impact varies across countries and domains.
- As no official robustness checks and justifications of the specification of the conditioning model are published, certain PISA results should be considered and used carefully, especially in the minor domains.

Chapter 3 'The effect of background variables and design choices on student achievement scores: A simulation study based on PISA 2012' further investigates the conditioning model, though focusing on different properties of the data used in the process. More specifically, I am interested in the way that the background variables are prepared for conditioning, whether it matters if all students are administered questions in all domains (something which is not done in PISA) and the impact if all students were to answer all background questions instead of a subset. To approach these research questions, a simulation study based on the PISA 2012 test design and data is conducted. This allows me to investigate questions that would not be otherwise possible. My key results are:

- The simulation study confirms that the conditioning model matters:
    - In general, bias decreases when conditioning is applied.
    - Yet, depending on the exact conditioning model specification, bias can remain present in distinct ways, especially in the minor domains.
    - Bias tends to be larger on average in the minor domains.
    - In this setting, one variable – booklet ID – introduces substantial bias, especially in reading.

- It is not important how the variables are prepared, i.e. whether or not a principal component analysis is conducted prior to conditioning in order to reduce the dimensionality of the background variables.
- Bias is much smaller for students who answer questions in all test domains.
- The impact of the background questionnaire design (whether or not all students were administered all questions) is negligible.

Chapter 4 ‘Group comparisons in PISA: What can go wrong along the way? A case study of differences in achievement by parental education in Germany aims to show a more comprehensive picture of potential sources of bias that can impact student achievement scores through different pathways. The following six statistical properties that can potentially influence PISA scores and group comparisons are identified and investigated:

1. Target population and sampling rates,
2. Survey non-response, i.e. school and student non-response rates as well as questionnaire return rates,
3. Item non-response,
4. Construction and coding of valid international scales,
5. Validity of students acting as proxy respondents for their parent(s),
6. Usage of background variables in the conditioning model.

In order to facilitate in-depth investigations of the six aspects, a case study of the impact on group comparisons is conducted: Highest parental education in Germany in PISA 2012. The official published PISA 2012 data, the more detailed German national version (available upon request) and the German Socio-Economic Panel (SOEP) data are used in this study. The key findings are:

- While not all of the areas investigated introduce substantial bias, some did.
- In Germany, the target population equals the complete population of 15-year-olds. The non-coverage rate of 2.6% is small and has little potential impact.

- While school and student response are comparatively high in Germany, questionnaire return rates are low. Overall, only 80% of the students and just above half of the parents returned their questionnaire.
    - Comparisons of PISA and SOEP data showed substantial differences between the distributions of highest parental education in the samples.
- While survey non-response was more frequent than item non-response for highest education, both occurred in the data.
    - Non-response is related to both highest parental education and student achievement and thus introduces bias into comparisons.
- Errors in coding the national qualifications onto the international scale for highest parental education are found. As a result, half of the low-educated parents in the published PISA dataset were incorrectly classified as such.
    - Achievement gaps between students with high- and low-educated parents were attenuated by 30 PISA points.
- Just above 50% of students and parents agreed on highest parental education.
    - Group comparisons differ by more than 25 points depending on whose response it is based on.
- Bias is introduced into group comparisons through the conditioning model when the background variables are measured with error. However, the magnitude of the bias is relatively small.

5.2 Overarching contribution

The three main chapters provide multiple important findings for different psychometric and statistical methods and properties using different approaches. However, they all have a common goal: to shed new light on the methodology and properties behind PISA, and how this then influences the PISA scores. Each individual study provides new detail from alternative – yet complementary – perspectives. Three common contributions emerge:

1. Adding to the literature

All three studies deal with parts of the PISA methodology that have received comparatively little attention so far. As a result, each chapter already adds to the literature on its own, but they also make a valuable contribution when taken together. As they highlight, it is important to investigate issues both from a detailed and a comprehensive perspective. In-depth investigations foster a deeper understanding of the workings behind a specific aspect of PISA. Analyses from a broader view can examine the cumulative impacts of different issues on the outcomes. This thesis therefore adds new knowledge to the literature about the conditioning model specification, different properties surrounding the conditioning model and more general statistical properties which affect the validity and comparability of PISA scores. This can be used to show a complete picture, highlighting how these different aspects are intertwined, and – in some situations – need to be considered simultaneously. Together, they demonstrate how PISA is complex, with issues arising across several different areas which may potentially introduce bias into student scores.

Revisiting the key literature of Chapters 2-4, I want to highlight how this thesis adds and relates to the existing literature in detail. Thereby, Chapter 2 and 3 both address the conditioning model in PISA and therefore mostly have joint key literature and contributions. While the fourth chapter also deals with methodological aspects in PISA, it adds to the literature in a different way.

The key background literature to Chapter 2 and 3 is the PISA 2012 technical report, which describes the technical details, i.e. how the PISA scores are computed among others. This thesis, thereby, adds to the literature by challenging some of the decision and characteristics, while also describing the procedure used to derive achievement scores in a clear and (hopefully) digestible way. The same cannot be said for the official documentation which spreads information across chapters, buries them/the details in text, appendices and technical language, if they are available at all. Furthermore, no justification or reasoning behind decisions are given which I highlight and question in those two chapters.

Other studies have also looked at the estimation of student achievement in PISA but with a different focus and aim. For those two chapters, the papers from Kreiner & Christensen (2014) and Rutkowski (2014) are the key previous studies. Kreiner & Christensen (2014) investigate the fit and validity of the PISA model as a whole. Thereby, they focus on one domain and the fit of the IRT model (but including conditioning to some degree). They found misfit and DIF which challenges the validity of the PISA results. The topic of appropriate IRT model choice model choice has also received attention from other researchers and points of views (e.g. Jerrim et al., 2018; Wuttke, 2007). While I also investigate the validity of the PISA scores, my focus is distinctly different – namely on the impact of the conditioning model, which they do not investigate. Rutkowski (2014) also focuses on the conditioning model using a simulation study and thereby showed how sensitive it is. But in comparison to my studies, while she motivates her study with the example of PISA, she uses a more general, constructed approach to highlight specific points. The joint contribution of the two chapters is to show that the selection of background variables in the conditioning model has an impact on the PISA scores and can bias results.

Chapter 3 has some additional interesting results regarding the (test) design of the cognitive assessment and the background questionnaire. One is the finding that asking students questions in all domains can help to reduce bias, especially in the minor domains. While it is a rather evident result, it adds an important point to the literature, as it can have a major impact on results and minor domains is a controverse unique feature of PISA. This finding is especially interesting in combination with the observation that it does not matter as much if the background questionnaire has a complete design or not. Thereby, there is no clear consensus in the literature: Adams et al. (2013) reach similar conclusions, whereas von Davier (2013) argues against (theoretically) against its use.

Revisiting the literature regarding the fourth chapter, this thesis also makes a valuable additional contribution to the literature. While it is common in other areas of surveys to regard potential source of bias in the light of the Total Survey Error framework, it is the exception in ILSAs. The only other study

which comprehensively discusses different statistical and psychometric sources of error in PISA is Schnepf (2018). She thereby holistically describes different aspects which can potentially bias PISA results, but often in a more theoretical way without any analysis. Chapter 4 is in line with Schnepf (2018) by acknowledging and agreeing with the different sources of error and the importance of investigating them. But it differs in its approach; one of the main contributions of Chapter 4 is the use of real-life data to highlight and quantify the impact by using a case study consistently. Indeed, the case study shows that PISA 2012 data should not be used for comparisons of highest parental education in Germany, because multiple aspects cause bias. Those aspects have been covered in the literature before but only as isolated issues. (e.g. Jerrim & Macmillan, 2015; L. Rutkowski, 2014; see Chapter 1.3 and 4.2 for more). The contribution is to show both that these issues persist in the data and how multiple aspects exist at once and how they are connected.

2. Raising awareness for potential impact

With this thesis, I do not only want to contribute to the highly specialised academic debate surrounding the estimation of PISA scores and results, but also to reach non-expert audiences. Thereby, I intend to foster an understanding of the way achievement scores are estimated and how and why they can be affected. While I state formulae and try to be formally correct, I also aim to use language and examples that make my findings easy to communicate. The analysis strategies (real-life data, simulation and case study) are chosen in order to give different perspectives on the workings and potential impact in clear ways. Food for thought and recommendations for different audiences consequently emerge (e.g. input for study makers about the implications of different study designs). Overall, an appeal for carefulness about the implications and limitations of PISA are raised for all consumers of this data – academic and non-academic.

3. Calling for transparency and scrutiny

Throughout the thesis, issues of transparency and replicability have surfaced. The lack of sufficiently detailed documentation – and published code – makes it hard for independent researchers to scrutinise certain aspects of the PISA

methodology. While the official technical documentation describes the basics and ideas behind PISA, many properties, computational details and decisions are just described as given without explanation, or simply not stated at all. In a first step towards more transparency and reproducibility, I have published my code to allow researchers to judge my research on their own and reuse it if they want (Zieger, 2021), yet the OECD does not currently do the same. Many properties, which can potentially influence the PISA scores and results, are presented without justification or robustness checks, which hinders the examination of data quality and validity, especially for non-specialists. Such investigations and robustness checks require large amounts of technical knowledge, computational effort and time. It is absolutely vital that threats to the validity and comparability of PISA scores and results, which are used worldwide, are transparently and comprehensibly communicated to researchers, policy-makers, stakeholders and the media. This thesis shows how those threats come into existence.

While those three overarching contributions are all valid and important, their relevance and application vary for different audiences. Looking at the whole picture, different recommendations emerge for different audiences – academia, test makers and the general public.

The key point for academia is the relevance and potential impact of different psychometric and statistical properties in PISA. While other aspects also influence and bias the scores, those topics should be overlooked or ignored. I showed that those aspects can have an important impact on PISA results. Some much-needed further research is described in the next subchapter. I also urge researchers to be more outspoken about and involved in the open science movement, especially when dealing with such complex situations.

Test makers can benefit from food for thought from this thesis. Some results show interesting leverage points (e.g. free up space by using incomplete background questionnaire designs for additional items) which should be considered carefully and kept in mind, if not implemented. Furthermore, this thesis clearly highlights the need for better communication on multiple levels

– better technical documentations for ‘expert audiences’ and more transparent and prominent descriptions what the data can and cannot do as well as an explanation of potential drawbacks and impact for ‘general audiences’.

My recommendation for the general public is to take everything with a pinch of salt and thinking critically before taking all the (media) reports at face value. Challenge the inferences drawn from PISA and whether this is the best data available. While it might not be perfect, it is the most comprehensive and best data in various situations. Yet, it should always be treated care and a healthy dose of doubt instead of as the pure truth.

5.3 Limitations

While I believe that this thesis makes valuable contributions to both the academic literature and general debates about PISA and its use, this research is also not without its limitations.

First, all studies are applied in a specific setting that either uses real-life data or is closely based on real-life data and design. This is not necessarily negative, as it allows for more detailed investigations into the setting and situations, but it does come with limitations. Most notably, the generalisability of findings outside of the specific setting investigation may be limited. In this thesis, all chapters have been based on PISA 2012 due to its interesting design and particularly rich data available. Yet, ILSAs differ in certain specific characteristics and change over time. As a result, findings may not be directly transferable to other settings (e.g. other ILSAs). They can nevertheless raise awareness of key aspects surrounding the methodology, potential biases and the associated implications for cross-national comparability.

Second, while I spent great time, effort and care to align the methodology in this thesis as closely as possible to the procedures used in PISA 2012, I cannot rule out some differences remaining. In Chapter 2, I tried to replicate the official PISA scores and found that the official technical documentation does not contain all computational details and relevant information. Although my replication of the PISA 2012 scores was very close to the reported values, the correlation was not perfect. As Chapter 3 is based on the same methods as

Chapter 2, the problems due to the lack of documentation also apply there. Additionally, I found in Chapter 4 that details (numbers and descriptions) in the technical reports do not always match between the national and international documentation. Overall, the lack of complete documentation and publicly accessible replication materials is an important – and underappreciated – limitation of PISA (and, indeed, potentially for other ILSAs as well).

Third, it was not always feasible to conduct the whole student achievement score process as implemented in PISA. In all three chapters, I did not compute some of the many steps myself – but rather used information published by the OECD in their technical reports (e.g. published estimates or formulae) as my starting point. For instance, I have not recomputed the scale transformation (a linear transformation) to allow my scores to be placed upon the PISA scale over time (as I have not undertaken cross-cycle comparisons). While this might have impacted absolute differences between official and self-computed scores, relative analyses, such as correlations and rankings, are not affected by such issues. Relatedly, one caveat in Chapter 3 is that it was not feasible to simulate the organisational structure of the PISA data (i.e. students nested within schools).

5.4 Future research

This thesis has generated new knowledge and insights into the process behind PISA scores and results. Yet while it started to shine new light on some aspects, much further work is needed. Indeed, this research did not only lead to new insights, but also to new research questions and ideas.

The need for future research into the achievement score computation clearly comes through. Chapter 2 and Chapter 3 showed the importance of the conditioning model and the properties surrounding it. However, little research is conducted on that matter. It would be interesting to compare this research based on PISA 2012 to similar new research based on other situations, such as other cycles or other ILSAs which come with a different set of specifications and properties. Since the establishment of PISA, the test designs and methodologies are constantly developing. It is therefore of great

importance that research keeps up to date, especially as new technologies – such as computer adaptive testing – are becoming increasingly used. Different ILSAs integrate these advancements into their methodologies in different ways. This thus adds new dimensions to student achievement estimation and conditioning which should be closely scrutinised.

The findings from Chapter 3 illustrate how bias could be substantially reduced if all students are administered questions in all domains. In contrast, the impact of the background questionnaire design was negligible. Yet most ILSAs currently use a complete background questionnaire design. As a result, it would be interesting to further examine different scenarios derived from these findings. For instance, the background questionnaire design from PISA 2012 could be adapted in other cycles and studies in order to include new topics in the student background questionnaire. If simulations and pilot studies confirm this, it could add additional value to the data.

Chapter 4 concerns itself with different properties and aspects in the broader picture of PISA which can potentially introduce bias into the scores and results. This draws on the total survey error framework. While I do not claim to capture the (complete) total survey error of PISA, I investigate multiple sources of bias, with little other research done on this topic for ILSAs. Further research on this matter would be hugely beneficial to gauge the quality and validity of ILSA data. Potential future research approaches could involve more detailed, currently not-published data from the OECD, such as information about non-participating schools and students. Another way to examine bias in the scores and results related to different statistical and methodological properties is to link the PISA data to national administrative (school and student) data.

A Appendices Chapter 2

A.1 Which countries participated to what extent in PISA 2012?

As explained in the main body of the chapter, 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 (Par). 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	Abbr.	n	PS	DRM	Par	ICT	EC	EB
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						
Luxembourg	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	

			Domain		Questionnaire			
Country	Abbr.	n	PS	DRM	Par	ICT	EC	EB
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				
Vietnam	VNM	4959						X

Note: Abbreviations: Abbr. = Abbreviation, n = sample size, PS = Problem Solving, DRM = Digital reading and mathematics, Par = Parental, ICT = Information and communication technology, EC = Educational career, EB = Easier booklets. 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 chapter, I do not consider them as administrating this domain.

A.2 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; OECD, 2014b). I try to illustrate the subtle differences using Figures A.1 and A.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 based upon a common dataset encompassing all countries.[46]

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[47] (see Figure A.1). Figure A.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 the paper-based core domain reading, science, and mathematics items. This is because 'CBA [computer-based assessment] reporting scale cannot influence the PISA paper-based assessment' (OECD, 2014b, 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

[46] The common sample exists of 500 students from each country, except for Liechtenstein, which were randomly selected.(OECD, 2014b, p. 233)
[47] The published item difficulties are used in my analyses.

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

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

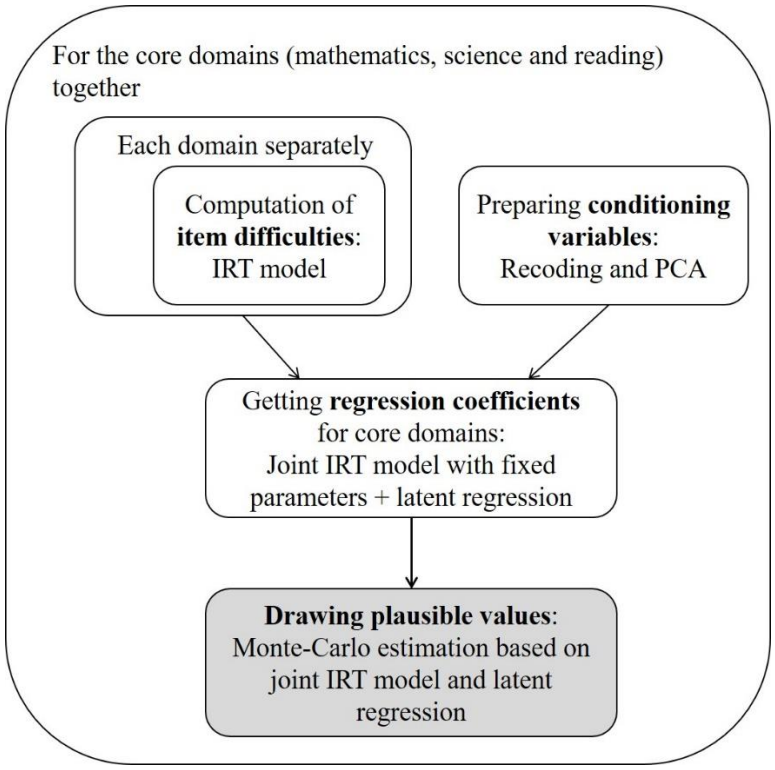
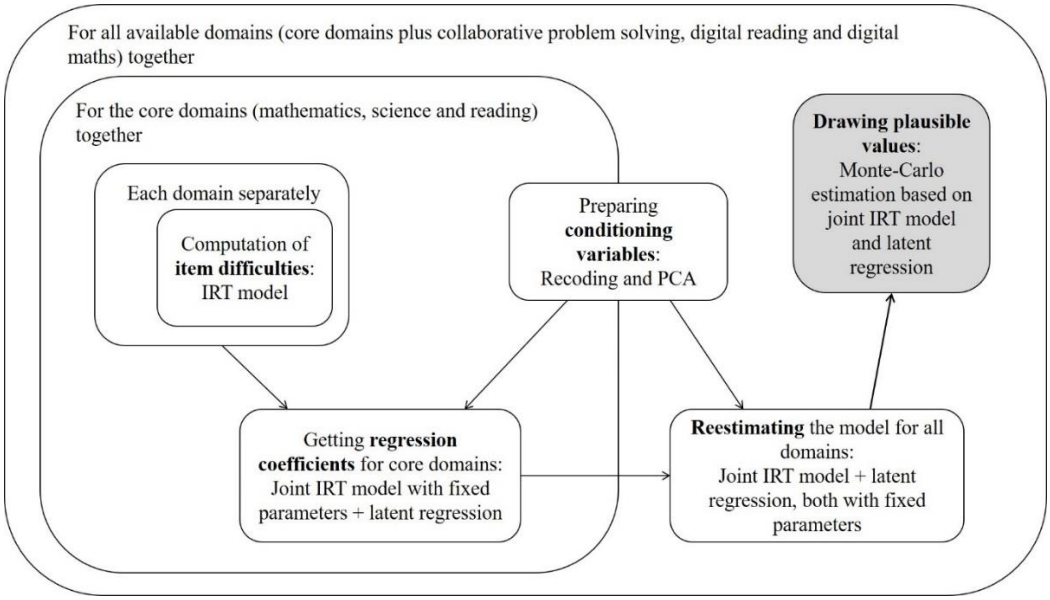


Figure A.2 Computation process of the plausible values, if the country administered additional domains (problem solving and/or digital reading and mathematics) to the three core domains



A.3 Computational details of the conducted analysis

This appendix attempts to make the computational procedures I have used as transparent as possible. All code used within my analysis is available from Zieger (2021). My empirical approach used the following steps:

0. Test data preparation. The already scored cognitive dataset 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, 2014b, pp. 231, 232) and if missing data was coded correctly (omitted and invalid treated as incorrect and not reached treated as missing; OECD, 2014b, pp. 233, 399).
1. Item difficulty estimation. As this is not the focus of this chapter and I do not want the conditioning model to be influenced by estimation

of own item difficulties. I therefore chose to use the published item difficulties within this analysis (Annex A; OECD, 2014b)[48].

2. Preparation of conditioning variables. The conditioning variables were computed from all available questionnaires, for each country and each assessment booklet. I used a two-stage process: recoding (stage 1) and pre-processing (stage 2). For the recoding and first pre-processing, I adhere to the recoding procedures as described in Annex B in the PISA 2012 technical report (OECD, 2014b, 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, 2014b, 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, I do not use polychoric correlations or other adaptions. In other words, I try to stay as close to the PISA technical report as possible (OECD, 2014b, p. 157). From this PCA, within each country I retained enough principal components to explain 95% of the variance in the data. This resulted in up to 153 principal components being extracted (and a minimum of 55) within each country, see Appendix A.8 for further details on the number of principal components. 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, 2014b, p. 157). As a consequence, I split the data of those countries into subsamples by alternatingly assigning strata to the new datasets starting with the largest strata.

[48] It is worth noticing that I believe that the step difficulty of item PM155Q03D is a typing error. I 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 I 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)[49]. I still experience computational difficulties in five countries (South Korea, Liechtenstein, Columbia, Shanghai (China), 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. I 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, 2014b, 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 chapter. Therefore, I use the formulas from the technical report (OECD, 2014b, pp. 253, 254). For the sake of convenience, I 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, I set seeds for the computation. I reran the analysis with different seeds but ended up with similar conclusions.

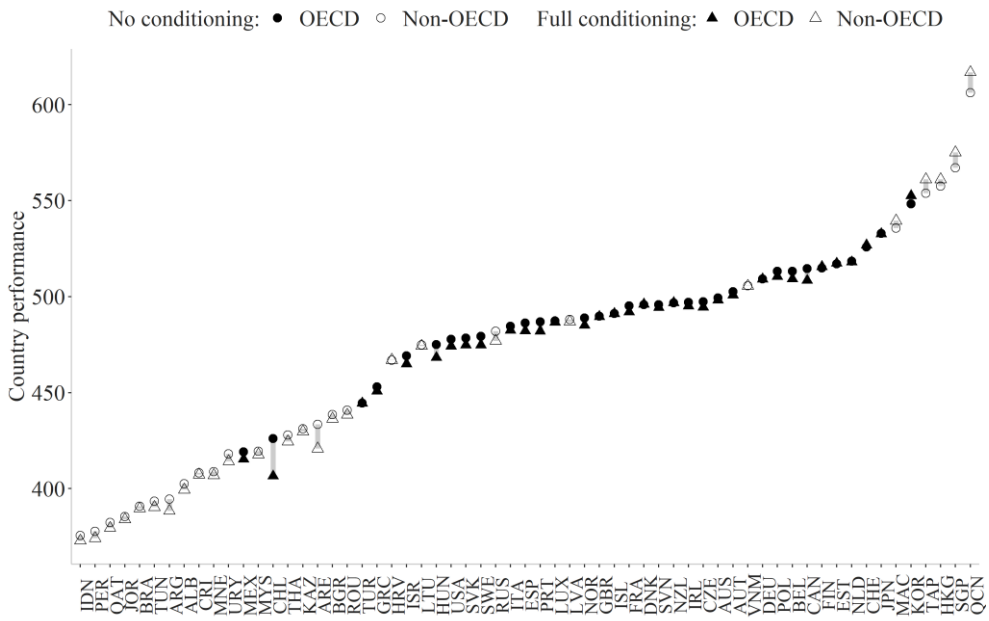
[49] This approach does have some limitations, however. For instance, it ignores the uncertainty in parameter estimates within the latent regression.

A.4 Mathematics: Domain specific analyses

A.4.1 Average scores

Figure A.3 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.

Figure A.3 Country average mathematics scores with and without conditioning



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

Table A.2 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 3 points from no conditioning (492 points) to full conditioning (489 points).

Table A.2 Variation in estimated average PISA mathematics scores by conditioning model specification in the OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
South Korea	548	557	555	-	554	561	554	553
Japan	533	532	530	530	533	531	530	533
Switzerland	526	527	528	527	527	527	527	527
Netherlands	518	518	518	519	518	518	518	518
Estonia	517	518	520	516	519	521	519	517
Canada	515	509	509	514	508	514	506	508
Finland	515	515	516	516	515	516	516	516
Belgium	513	513	515	509	511	510	509	509
Poland	513	508	512	510	511	510	510	511
Germany	509	510	514	509	512	509	510	509
Austria	503	503	505	501	501	500	501	501
Australia	499	500	500	500	499	498	498	498
Czech Republic	497	494	496	495	495	494	495	494
Ireland	497	501	499	495	497	495	491	495
New Zealand	497	497	497	497	497	497	497	497
Denmark	496	501	499	495	499	496	498	496
Slovenia	496	492	494	495	494	495	494	494
France	495	491	496	493	496	492	493	492
Iceland	491	491	491	491	491	491	491	491
United Kingdom	490	490	489	491	490	491	490	490
Norway	489	490	492	492	492	486	488	485
Luxembourg	487	487	487	487	487	487	487	487
Portugal	487	484	483	482	484	482	481	482
Spain	486	483	484	484	482	483	484	482
Italy	485	483	483	483	483	483	483	483
Sweden	479	476	479	482	477	480	476	475
Slovak Republic	478	474	480	482	474	474	479	475
USA	478	471	473	474	473	473	474	474
Hungary	475	469	470	468	469	468	468	468
Israel	469	467	463	464	464	467	462	465
Greece	453	452	451	452	451	452	451	451
Turkey	445	445	444	444	445	445	444	444
Chile	426	399	402	403	400	408	408	407
Mexico	419	416	416	416	416	416	416	415
OECD average	**492**	**490**	**491**	**488**	**490**	**490**	**490**	**489**
OECD median	**496**	**492**	**495**	**493**	**495**	**493**	**492**	**493**
Cor with M0	**1.00**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Cor with M7	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Note: 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 A.3. M0 = no conditioning; M1–M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors, 2: individual direct regressors, 3: indirect regressors, 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 A.3 Variation in estimated average PISA mathematics scores by conditioning model specification in the non-OECD countries

Country	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
Albania	403	400	399	400	399	400	399	399
United Arab Emirates	434	413	412	418	415	415	419	421
Argentina	394	389	390	390	388	389	389	388
Bulgaria	439	437	436	437	436	438	436	436
Brazil	391	390	388	390	389	390	389	390
Colombia	386	-	384	371	367	365	373	367
Costa Rica	408	408	407	407	408	408	407	407
Hong Kong (China)	558	563	565	564	563	562	563	561
Croatia	467	467	466	467	467	467	467	467
Indonesia	376	374	373	373	373	374	373	373
Jordan	385	384	383	384	384	384	384	384
Kazakhstan	431	431	429	430	430	431	429	430
Liechtenstein	532	531	531	533	531	-	532	-
Lithuania	475	474	474	474	474	475	474	474
Latvia	488	488	487	487	487	487	487	487
Macao (China)	536	540	540	540	540	540	540	539
Montenegro	409	408	406	407	407	408	406	407
Malaysia	419	418	417	418	417	418	418	418
Peru	378	375	374	375	374	375	374	374
Qatar	382	380	379	380	379	380	380	379
Shanghai (China)	606	605	-	618	-	613	619	617
Romania	441	439	439	440	438	440	438	438
Russian Federation	482	477	476	476	477	477	476	477
Singapore	567	576	577	576	576	575	576	575
Serbia	447	447	447	-	447	-	-	-
Chinese Taipei	554	563	564	560	563	560	562	561
Thailand	428	425	424	425	424	425	424	424
Tunisia	393	392	390	392	390	392	390	390
Uruguay	418	416	415	416	414	415	415	414
Vietnam	506	506	506	506	506	507	506	506

Note: 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. Columbia, Liechtenstein, Shanghai (China) and Serbia are missing the scores due to computational difficulties.

A.4.2 Inequality in PISA scores

I am not only interested in the country average scores, but also in inequality measures. The difference between the 90^{th} and 10^{th} percentile is an inequality measure for spread and displayed in Table A.4. 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 0.69 and 0.8. Furthermore, the average OECD percentile difference experiences a sharp rise from 215 to somewhere between 247 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 are also relative changes between the different specifications (see differing colour patterns). The correlations between the conditioning specifications with full conditioning range between 0.78 and 0.97 with especially high correlations ($r > 0.9$) for all specifications except for school direct regressors only.

Table A.4 Estimates of inequality (90^{th}–10^{th} percentile difference) in PISA mathematics scores across countries by specification of the conditioning model in the OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	158	183	186	183	185	183	185	185
Estonia	190	247	235	244	241	230	242	244
Chile	194	190	197	194	192	229	224	211
Denmark	196	216	253	253	226	251	224	245
Ireland	196	254	243	253	252	252	210	248
Finland	197	216	213	217	215	217	217	217
Spain	201	223	227	225	228	227	227	229
Greece	204	229	227	226	228	227	227	228
Canada	205	242	240	240	244	233	249	248
Norway	210	269	271	272	271	269	272	265
Slovenia	211	279	232	268	258	273	238	256
Sweden	212	172	265	218	268	232	259	257
Iceland	213	235	235	234	237	235	236	237
USA	215	263	268	271	270	270	271	270
Austria	216	272	278	269	271	266	270	266
United Kingdom	216	239	240	238	240	239	241	240
Italy	216	241	241	240	242	241	241	242
Switzerland	218	244	243	242	245	244	243	245
Poland	218	237	275	276	275	275	277	275
Japan	220	234	233	232	241	237	236	235
Netherlands	220	241	242	241	243	243	242	244
Turkey	220	238	237	240	237	238	240	238
Hungary	221	274	275	274	273	273	274	269
Portugal	221	248	282	282	273	280	281	276
Luxembourg	224	245	249	245	249	245	248	248
Australia	226	280	279	279	276	256	275	253
Czech Republic	227	247	249	247	248	246	250	249
Germany	228	291	291	291	293	287	290	285
France	228	283	293	293	288	278	293	270
New Zealand	233	256	257	256	258	257	258	258
Belgium	241	304	304	303	303	298	302	299
Slovak Republic	242	270	309	283	284	281	305	282
Israel	244	298	291	305	300	296	290	291
OECD average	**215**	**247**	**253**	**253**	**253**	**252**	**253**	**252**
OECD median	**216**	**245**	**249**	**247**	**249**	**246**	**248**	**248**
Cor with M0	**1.00**	**0.69**	**0.79**	**0.75**	**0.80**	**0.77**	**0.82**	**0.79**
Cor with M7	**0.79**	**0.78**	**0.95**	**0.94**	**0.97**	**0.95**	**0.92**	**1.00**

Note: Figures illustrate how the difference between the 90^{th} and 10^{th} 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 A.5. 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Table A.5 Estimates of inequality (90^{th}–10^{th} percentile difference) in PISA mathematics scores across countries by specification of the conditioning model in the non-OECD countries.

Country	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
Albania	184	212	212	212	213	212	213	213
United Arab Emirates	209	254	236	259	254	258	257	255
Argentina	166	190	193	191	192	190	194	193
Bulgaria	222	244	244	243	244	244	244	245
Brazil	174	197	200	198	198	196	199	198
Costa Rice	143	168	172	168	172	168	170	170
Hong Kong (China)	217	284	276	282	282	284	280	282
Croatia	207	228	222	230	227	231	230	231
Indonesia	150	176	179	178	178	177	179	178
Jordan	169	195	198	195	197	194	197	197
Kazakhstan	156	180	183	181	182	180	182	181
Lithuania	209	232	233	232	232	232	233	232
Latvia	188	211	214	212	213	210	214	215
Macao (China)	214	256	259	263	261	264	265	265
Montenegro	187	210	207	210	210	211	208	211
Malaysia	185	207	210	209	209	208	209	209
Peru	180	207	208	206	209	207	208	208
Qatar	228	251	253	249	253	249	251	252
Romania	188	210	214	212	211	210	213	212
Russian Federation	201	217	214	221	223	223	219	226
Singapore	249	315	314	316	314	318	317	316
Chinese Taipei	276	347	356	347	346	348	359	354
Thailand	194	208	215	209	211	207	212	208
Tunisia	176	197	204	200	200	197	203	199
Uruguay	199	224	224	223	224	223	223	224
Vietnam	196	221	220	222	222	222	222	222

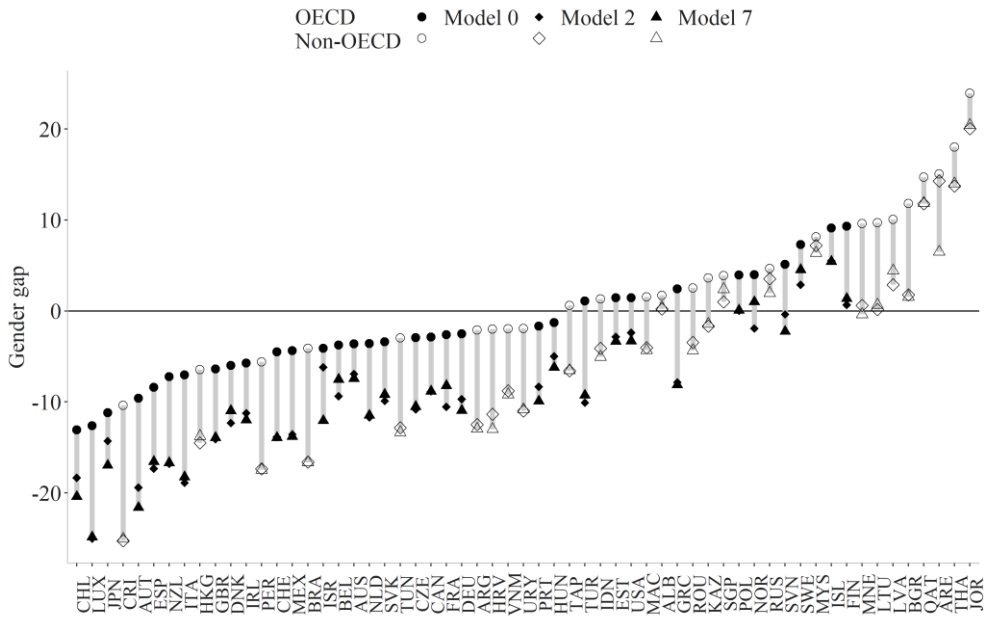
Note: Figures illustrate how the difference between the 90^{th} and 10^{th} percentile of PISA mathematic scores changes depending upon the specification of the conditioning model. M0 = no conditioning; M1–M6 correspond to conditioning with different subsets of conditioning variables (1: school direct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

A.4.3 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 chapter, I 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 principal component analysis).

Figure A.4 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 be found (0 points on average), while boys perform 6 to 7 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 1 point with full conditioning. Overall, the diamonds (individual direct regressors only: M2) and triangles (full conditioning: M7) mostly sit on top of each other and are distinct from the circles (no conditioning: 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.

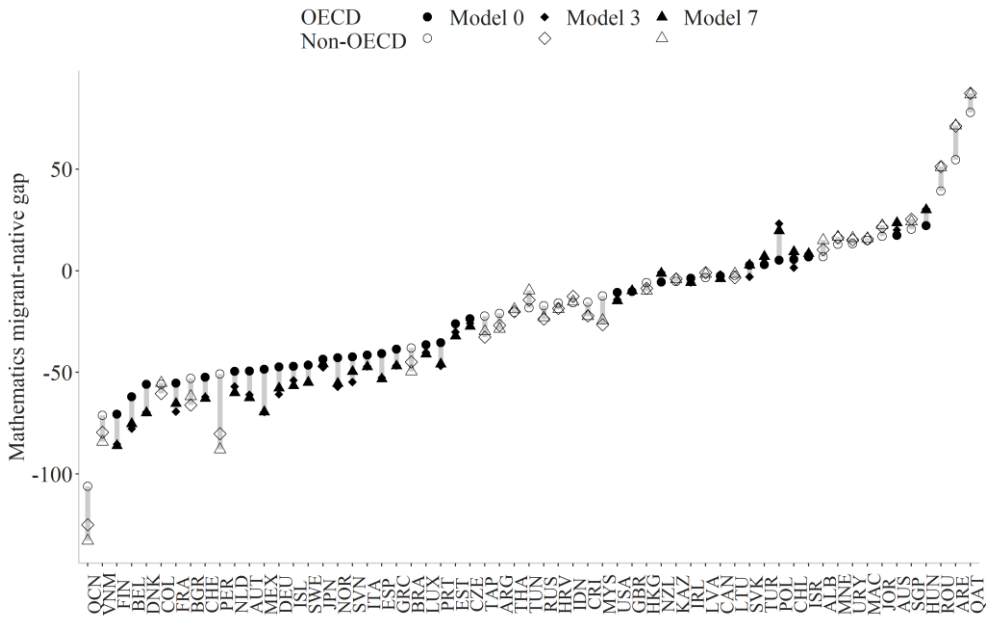
Figure A.4 Country mathematics gender gap without conditioning (M0), just with individual direct regressor including gender (M2) and with full conditioning (M7)



Note: 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. The two boxes highlight the examples given in the main text for substantial differences between the gender gap for M2 and M7.

The achievement difference between migrant and native students in mathematics can be seen in Figure A.5. Overall, the three symbols – circle (no conditioning: M0), diamond (indirect regressors only: M3) and triangle (full conditioning: M7) – sit roughly on top of each other for smaller gaps (around zero) with a slight increase in magnitude for M3 and M7 if a gap is present (i.e. gap is not around zero). The average migrant-native gap drops from -20 points (M0) to -24 (M3) and -24 points (M7), but the gaps in the countries themselves cover a rather big range from -133 points (M7 in Shanghai) to +87 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 as long as migrant status is included in the conditioning model.

Figure A.5 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)



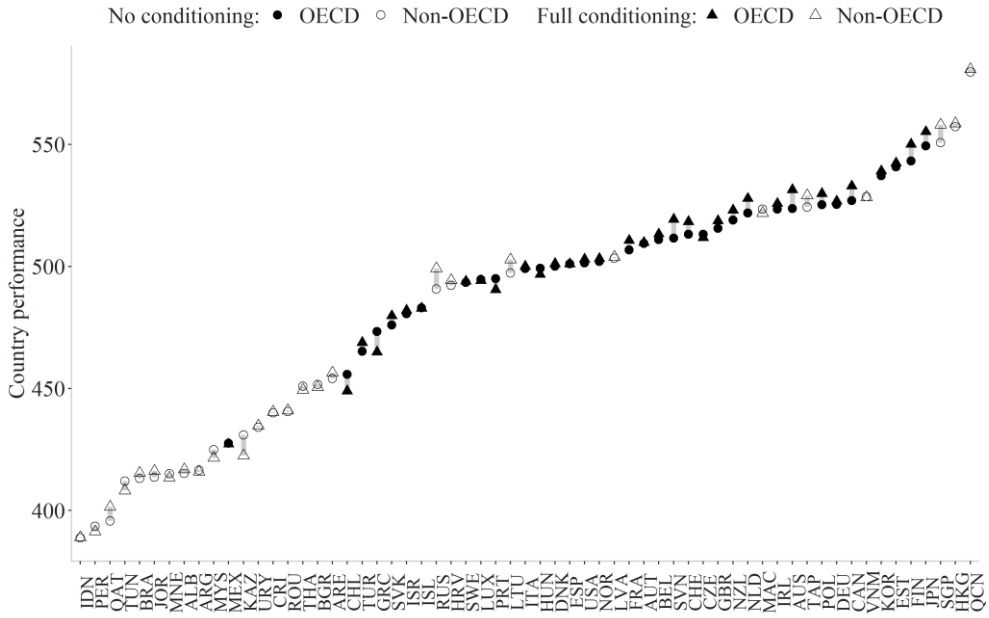
Note: Circles provide estimates without conditioning, diamonds for conditioning only with individual indirect regressors and triangles for full conditioning. Solid markers denote OECD countries and hollow markers non-OECD countries.

A.5 Science: Domain specific analyses

A.5.1 Average scores

Figure A.6 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 2 points when full conditioning is applied, but there is no common direction. At the extremes, Russia experiences an increase of 8 points, while Greece experiences a decrease of -8 points. Yet the ranking of countries remains roughly the same.

Figure A.6 Country average science scores with and without conditioning



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

This is stressed by Table A.6, 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) are 0.99 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 (505 points) to full conditioning (507 points).

Table A.6 Variation in estimated average PISA science scores by conditioning model specification in the OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Japan	549	550	547	552	546	551	549	555
Finland	543	548	546	547	550	548	548	550
Estonia	541	541	545	535	544	537	543	542
South Korea	537	535	541	-	541	532	539	539
Canada	527	530	534	520	533	525	529	533
Germany	525	524	528	523	525	524	526	527
Poland	525	533	531	528	531	528	531	530
Australia	524	527	528	525	529	531	530	531
Ireland	524	523	525	523	522	523	532	526
Netherlands	522	522	527	522	527	522	529	528
New Zealand	519	519	522	519	523	519	523	523
United Kingdom	516	517	518	516	519	516	518	519
Switzerland	513	517	517	515	519	517	517	518
Czech Republic	513	510	513	512	512	510	513	512
Slovenia	512	510	523	510	519	508	522	519
Belgium	511	510	512	512	510	513	512	513
Austria	509	510	513	509	510	508	510	510
France	507	504	510	513	503	506	515	511
Norway	502	501	501	502	502	502	501	503
Spain	501	501	501	500	500	500	501	501
USA	501	508	504	502	504	500	503	503
Denmark	500	499	492	491	500	498	499	501
Hungary	499	494	500	496	494	494	496	497
Italy	499	497	502	497	502	497	500	500
Luxembourg	495	494	495	494	495	494	494	494
Portugal	495	495	490	488	498	488	489	490
Sweden	493	511	498	499	498	498	497	494
Iceland	483	482	483	481	483	481	483	483
Israel	481	473	489	475	483	474	489	482
Slovak Republic	476	479	478	468	478	476	476	480
Greece	473	472	464	472	466	472	465	465
Turkey	465	464	469	465	469	464	469	469
Chile	456	448	452	452	449	447	449	449
Mexico	428	426	427	425	427	425	427	427
OECD average	**505**	**505**	**507**	**503**	**506**	**504**	**507**	**507**
OECD median	**508**	**510**	**511**	**509**	**507**	**507**	**511**	**511**
Cor with M0	**1.00**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Cor with M7	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**1.00**	**1.00**

Note: 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 A.7. 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 regressors, 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 A.7 Variation in estimated average PISA science scores by conditioning model specification in the non-OECD countries

Country	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
Albania	415	413	417	412	417	413	417	417
United Arab Emirates	454	459	460	457	459	457	457	456
Argentina	417	414	416	413	416	414	416	416
Bulgaria	452	449	452	450	452	450	451	450
Brazil	413	413	415	412	416	413	414	415
Colombia	420	-	426	420	417	414	421	417
Costa Rica	440	439	440	438	440	438	440	440
Hong Kong (China)	557	559	560	559	560	557	559	559
Croatia	492	491	496	491	495	491	494	494
Indonesia	389	386	388	387	389	386	389	389
Jordan	414	413	415	411	416	412	415	416
Kazakhstan	431	428	430	429	420	428	430	423
Liechtenstein	528	529	530	532	531	-	530	-
Lithuania	497	496	503	496	503	496	502	503
Latvia	503	502	503	502	503	502	504	504
Macao (China)	523	523	524	523	523	522	523	522
Montenegro	415	413	415	412	414	412	413	413
Malaysia	425	423	421	423	422	423	419	422
Peru	394	390	394	390	393	389	392	391
Qatar	396	393	401	393	401	393	402	402
Shanghai (China)	580	582	-	577	-	575	583	581
Romania	440	439	442	439	441	439	441	441
Russian Federation	491	502	507	508	500	501	506	499
Singapore	551	557	555	555	557	559	555	558
Serbia	449	448	451	-	450	-	-	-
Chinese Taipei	524	530	530	520	528	524	528	529
Thailand	451	448	451	448	451	448	451	449
Tunisia	412	409	408	409	408	409	409	408
Uruguay	434	433	436	431	436	431	435	435
Vietnam	529	529	529	529	528	529	528	528

Note: 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. Columbia, Liechtenstein, Shanghai (China) and Serbia are missing the scores due to computational difficulties.

A.5.2 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 A.8, 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 Sweden, which was in the middle-to-bottom category (high inequality) for some specifications (M0, M1, M3 and M5) end up in the top category for others (M2, M4 and M6). Percentile differences in science depend on the exact specification of the conditioning model. This is reflected in the correlations, which were especially low between no conditioning and the other specifications $(0.36 \leq r \leq 0.66)$, but also not consistently high across the other specifications with values below 0.9 in some cases.

Table A.8 Estimates of inequality (90^{th}–10^{th} percentile difference) in PISA science scores across countries by specification of the conditioning model in the OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Mexico	139	164	164	164	164	165	165	165
Estonia	179	101	94	124	103	121	101	110
Chile	183	139	144	142	144	162	164	158
Turkey	183	207	202	205	206	207	206	209
Spain	187	212	211	213	213	214	212	214
Greece	196	223	224	224	226	225	225	228
Poland	197	176	173	174	179	180	178	186
Canada	201	150	126	142	153	161	157	177
Czech Republic	204	223	220	221	219	225	220	221
Switzerland	205	229	230	228	230	230	229	230
Hungary	205	183	186	187	184	185	188	194
Portugal	205	197	171	187	200	198	177	203
Ireland	207	145	138	163	151	168	171	182
Denmark	208	116	145	161	132	172	128	176
Italy	208	234	234	234	235	236	234	236
Austria	211	161	135	197	191	206	194	205
Slovenia	211	196	193	179	210	198	200	211
USA	213	186	173	170	179	183	176	184
Finland	214	235	234	232	232	232	229	230
Japan	214	227	227	226	239	231	227	273
France	219	185	147	169	174	205	177	219
Sweden	220	309	135	239	154	236	185	218
Germany	221	168	143	172	169	201	179	203
Netherlands	221	241	240	241	237	243	240	240
Norway	221	162	162	164	163	184	171	209
Iceland	222	246	243	250	244	251	248	248
Belgium	225	177	174	189	181	196	190	200
United Kingdom	226	251	250	249	251	251	249	251
Australia	231	206	197	204	215	223	212	224
Slovak Republic	235	223	166	202	222	235	170	235
Luxembourg	237	260	259	261	259	262	261	262
New Zealand	238	260	258	260	257	260	260	260
Israel	239	230	230	226	240	233	237	249
OECD average	**210**	**201**	**189**	**200**	**199**	**208**	**199**	**212**
OECD median	**211**	**206**	**186**	**202**	**206**	**207**	**194**	**214**
Cor with M0	**1.00**	**0.48**	**0.36**	**0.50**	**0.48**	**0.63**	**0.46**	**0.66**
Cor with M7	**0.66**	**0.83**	**0.82**	**0.89**	**0.91**	**0.95**	**0.89**	**1.00**

Note: Figures illustrate how the difference between the 90^{th} and 10^{th} 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 A.9. 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

Table A.9 Estimates of inequality (90^{th}–10^{th} percentile difference) in PISA science scores across countries by specification of the conditioning model in the non-OECD countries

Country	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
Albania	187	220	222	219	223	220	222	223
United Arab Emirates	209	175	166	196	178	182	201	201
Argentina	176	207	208	207	210	208	210	211
Bulgaria	235	260	262	263	263	261	263	263
Brazil	162	186	187	185	187	187	187	187
Costa Rice	135	162	163	161	165	164	163	165
Hong Kong (China)	181	103	95	96	110	113	107	123
Croatia	196	220	216	219	218	221	218	220
Indonesia	136	162	164	163	164	164	164	166
Jordan	174	200	204	206	203	204	206	204
Kazakhstan	155	185	185	185	188	185	187	185
Lithuania	197	221	223	220	223	221	223	223
Latvia	176	201	198	199	199	204	199	202
Macao (China)	177	102	99	110	107	117	112	123
Montenegro	189	214	216	213	215	213	214	214
Malaysia	173	195	198	197	199	199	200	201
Peru	151	178	177	178	180	180	179	181
Qatar	238	263	261	263	262	263	261	262
Romania	177	202	203	203	205	203	204	205
Russian Federation	188	177	166	170	186	186	174	188
Singapore	239	170	165	168	169	178	168	175
Chinese Taipei	195	138	131	155	135	156	133	136
Thailand	170	186	190	189	187	188	190	188
Tunisia	168	197	199	198	200	197	200	200
Uruguay	198	225	228	230	237	231	230	231
Vietnam	168	190	184	191	188	191	188	189

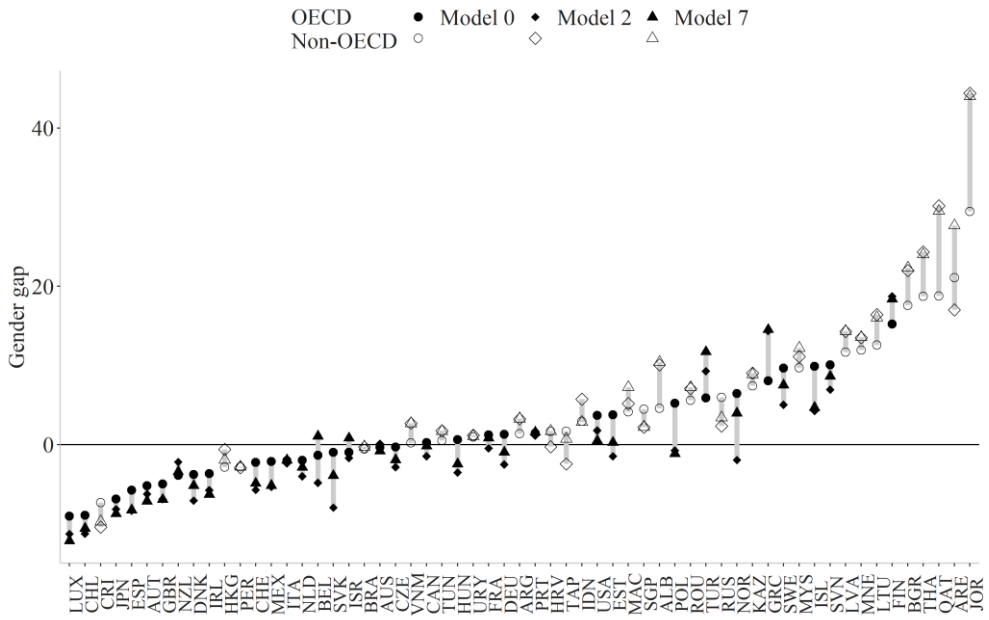
Note: Figures illustrate how the difference between the 90^{th} and 10^{th} percentile of PISA science scores changes depending upon the specification of the conditioning model. M0 = no conditioning; M1–M6 correspond to conditioning with different subsets of conditioning variables (1: school direct regressors,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

A.5.3 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 chapter, I 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 principal component analysis).

Figure A.7 shows that the conditioning model specifications have rather little influence on the gender gap in science in comparison to the gender gaps in mathematics and reading. In most countries, the two models including gender (M2 – diamond and M7 – triangle) sit close to each other, which means that the gender gap is robust as long as gender is included in the model. For some of 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 (e.g. absolute value always increases when conditioning is applied). For other, where the gender gap is located around 0, no difference between all three symbols can be seen. If there is no gender in the sample to begin with, it cannot be attenuated. As a result, the average gender gap stays the same on average between no conditioning (3 points) and full conditioning (3 points), even though single countries experience changes. Yet, some rank changes still occur, because not all countries are sensitive to the specification of the conditioning model and the direction of the impact varies.

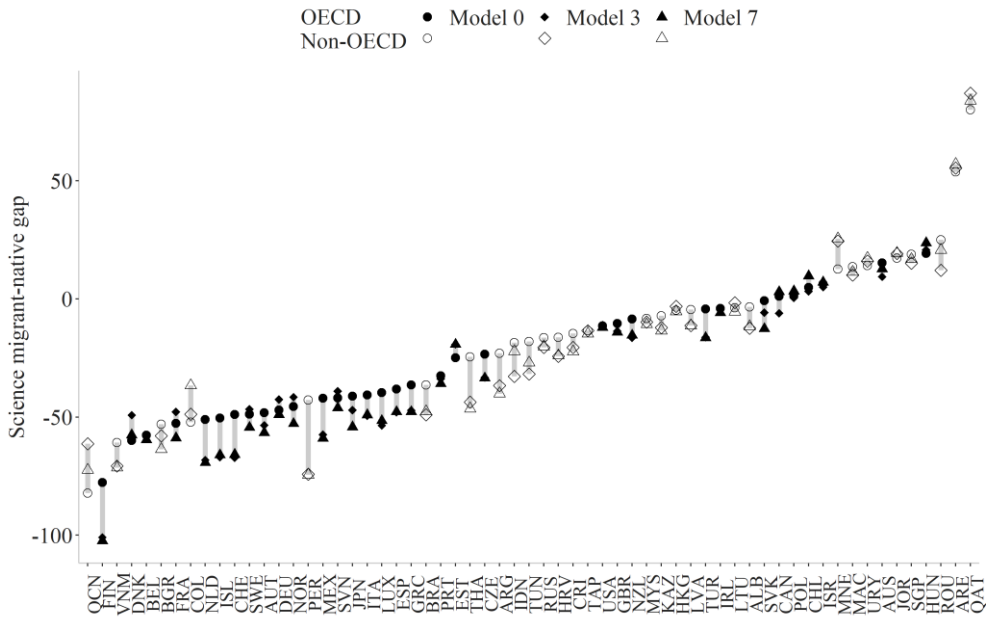
Figure A.7 Country science gender gap without conditioning (M0), just with individual direct regressor (incl. gender) in conditioning (M2) and with full conditioning (M7)



Note: 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 A.8 displays the gaps in science achievement for another grouping variable – migrant status (native versus migrant students). Similar to mathematics, the three symbols mostly are roughly in the same area if the gap is small or they increase in magnitude as soon as conditioning is applied. Thereby, the M3 and M7 are situated next to each other. While there are some exceptions, it seems like the migrant-native gaps are robust to the specification as soon as migrant status is included in the model.

Figure A.8 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)



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

A.6 Reading: Non-OECD specific tables

Table A.10 Variation in estimated average PISA reading scores by conditioning model specification in the non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Albania	409	406	409	407	407	406	409	406
United Arab Emirates	444	477	484	469	473	474	467	462
Argentina	406	401	402	402	401	401	402	402
Bulgaria	441	438	436	437	437	437	436	436
Brazil	414	414	411	413	413	414	413	414
Colombia	421	-	461	452	448	448	453	447
Costa Rica	448	446	447	445	446	445	447	447
Hong Kong (China)	542	533	528	531	526	536	526	529
Croatia	483	482	462	482	467	482	468	468
Indonesia	399	396	397	397	396	396	396	396
Jordan	405	402	398	402	398	401	399	399
Kazakhstan	395	392	386	392	381	392	386	385
Liechtenstein	513	514	504	504	503	-	501	-
Lithuania	475	474	468	474	468	474	468	467
Latvia	490	488	478	488	477	487	478	477
Macao (China)	509	493	490	491	490	493	490	491
Montenegro	425	424	426	423	426	423	424	425
Malaysia	403	401	404	400	402	400	402	400
Peru	401	398	401	398	400	397	401	400
Qatar	398	394	391	395	391	394	392	392
Shanghai (China)	565	587	-	581	-	575	560	558
Romania	434	433	435	433	435	432	435	435
Russian Federation	477	478	472	475	470	479	471	476
Singapore	538	532	524	530	525	534	522	526
Serbia	449	448	447	-	447	-	-	-
Chinese Taipei	520	514	511	528	509	524	507	508
Thailand	444	441	436	442	436	441	436	434
Tunisia	414	412	416	412	416	412	416	416
Uruguay	426	423	423	422	419	421	423	423
Vietnam	505	504	491	504	492	504	492	493

Note: 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning. Columbia, Liechtenstein, Shanghai (China) and Serbia are missing the scores due to computational difficulties.

Table A.11 Estimates of inequality (90^{th}–10^{th} percentile difference) in PISA reading scores across countries by specification of the conditioning model in the non-OECD countries

Country	M0	M1	M2	M3	M4	M5	M6	M7
Albania	217	258	264	257	266	259	264	268
United Arab Emirates	214	168	164	185	171	172	192	202
Argentina	200	233	237	236	237	236	239	240
Bulgaria	273	309	310	310	311	311	310	310
Brazil	184	216	224	216	224	217	223	223
Costa Rice	149	182	188	181	192	180	190	190
Hong Kong (China)	192	201	202	202	206	201	210	211
Croatia	193	219	240	216	228	215	224	223
Indonesia	155	188	198	188	198	189	198	198
Jordan	195	230	234	228	235	230	234	234
Kazakhstan	156	184	195	185	193	186	195	192
Lithuania	197	222	228	223	228	222	229	230
Latvia	182	210	214	208	217	210	217	218
Macao (China)	181	208	208	210	211	210	212	212
Montenegro	203	237	236	237	237	238	239	239
Malaysia	184	213	215	217	215	216	217	217
Peru	194	226	227	229	229	229	231	232
Qatar	251	286	284	284	286	285	285	286
Romania	202	234	237	232	238	234	237	239
Russian Federation	201	271	275	280	274	272	279	273
Singapore	232	220	206	213	218	240	215	230
Chinese Taipei	212	182	177	139	180	176	182	186
Thailand	173	196	203	201	204	199	207	201
Tunisia	190	227	229	226	232	227	229	232
Uruguay	204	239	238	238	248	240	240	243
Vietnam	161	183	189	183	187	180	188	186

Note: 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 regressors, 4: all direct regressors, 5: school direct and indirect regressors, 6: individual direct and indirect regressors); M7 = full conditioning.

A.7 How does the migrant-native gap in reading scores change when migrant status is used as a direct (rather than indirect) regressor?

Figure A.9 highlights the differences in the migrant-native gap in reading scores between three separate models:

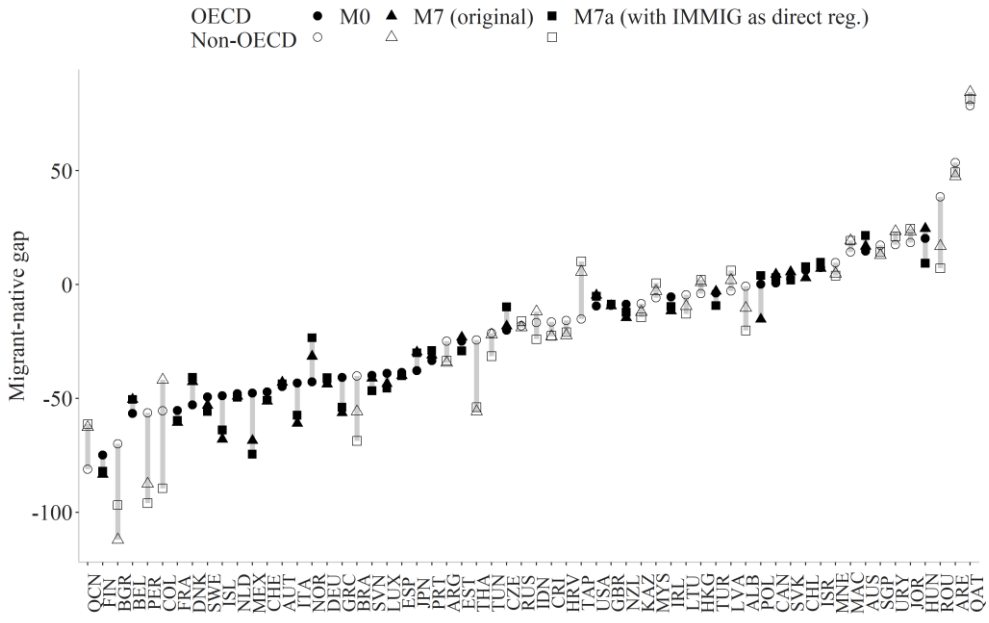
(i) No conditioning (M0).

(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 substantially 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 a few countries, such as Bulgaria (M7 = -112; M7a = -97) and Colombia (M7 = -42; M7a = -89), there is an appreciable change in at least the magnitude of the immigrant-native gap. These are, however, the exceptions rather than the rule and generally connected to very small percentages of migrants in the country. 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.

Figure A.9 Country reading gap between migrant and native students without conditioning (M0), with original model M7 and altered model M7a (migrant status included as direct regressor).



Note: Circles provide estimates without conditioning, triangles for full conditioning and squares for the altered full conditioning model with IMMIG included in the direct regressors. Solid markers denote OECD countries and hollow markers non-OECD countries.

A.8 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 (Booklets 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. I am therefore interested in how many principal components are retained with the design used in PISA 2012 and how the number of indirect regressors changes if I look at the separate booklet questionnaires only without missing by design and not complete rotated design with missing by design (see Table A.12). In the computations for this chapter, the principal components based on the design in PISA 2012 are used, see column 'All'. The maximum of retained principal components is 153 in Italy and the minimum is 55 in Liechtenstein. Overall, the number of principal components vary between countries and booklets, but while the single booklets have considerably fewer missing data, the number of retained principal components stays roughly the same in each country. The maximal number 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 lie in a plausible range with no surprising outliers anywhere.

Table A.12 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**
Australia	103	91	105	102
Austria	115	104	122	116
Belgium	145	137	148	144
Canada	102	92	103	100
Switzerland	109	100	110	103
Chile	130	117	132	131
Czech Republic	95	85	100	91
Germany	110	102	112	124
Denmark	120	111	122	120
Spain	106	94	108	101
Estonia	94	82	95	93
Finland	113	106	118	114
France	82	73	86	85
United Kingdom	84	76	85	81
Greece	102	93	110	103
Hungary	133	124	139	137
Ireland	115	103	118	113
Iceland	78	80	87	89
Israel	85	73	87	91
Italy	153	146	157	149
Japan	80	69	82	85
South Korea	142	131	146	144
Luxembourg	116	109	117	114
Mexico	136	130	140	133
The Netherlands	90	83	93	92
Norway	90	82	90	92
New Zealand	97	94	101	100
Poland	91	84	100	90
Portugal	150	142	152	155
Slovak Republic	120	113	129	123
Slovenia	116	108	119	118
Sweden	94	88	97	97
Turkey	100	92	104	100
United States of America	77	69	82	78
Non-OECD countries:				
Albania	69	60	78	81
United Arab Emirates	86	78	90	87
Argentina	95	88	100	99
Bulgaria	84	76	85	85

Country	All	Booklet A	Booklet B	Booklet C
Brazil	89	78	90	89
Colombia	89	76	91	94
Costa Rica	102	94	106	102
Hong Kong (China)	135	122	136	143
Croatia	137	129	139	134
Indonesia	90	84	92	88
Jordan	102	97	108	104
Kazakhstan	84	77	88	80
Liechtenstein	55	51	53	82
Lithuania	81	73	86	79
Latvia	121	106	120	119
Macao (China)	139	132	146	142
Montenegro	83	76	87	85
Malaysia	82	72	84	82
Peru	88	79	91	92
Qatar	83	72	85	88
Shanghai (China)	95	84	98	97
Romania	86	78	91	84
Russian Federation	102	90	104	99
Singapore	110	100	116	112
Serbia	118	110	118	120
Chinese Taipei	96	88	102	95
Thailand	86	77	88	80
Tunisia	83	76	87	88
Uruguay	109	99	110	111
Vietnam	81	76	89	83

Note: Column ‘All’: The principal components are computed based on the background questionnaire data of the whole sample. Columns ‘Booklet A’, ‘Booklet B’, and ‘Booklet C’: The principal components are computed based on a subset of background questionnaire data with only the students who were administered booklet A, B or C respectively. In all situations, enough principal components were retained to explain 95% of the variance.

B Appendices Chapter 3

B.1 Does a single component of the direct regressors introduce bias to the results?

In the main body of Chapter 3, I found that bias increased in average reading scores as soon as direct regressors were added to the conditioning model specification. As a result, it is of interest to see if a single component of the direct regressors triggered this or whether it is an interplay of multiple factors. The direct regressors consist of four main components – booklet IDs, gender, parental socio-economic status (ISEI) and grade – which were prepared as in PISA 2012. In order to disentangle the impact that these variables have, I analysed the following six models of which four models included only one of the components of the individual regressors in the conditioning model:

V0. No conditioning variables (i.e. no conditioning model at all)

V1. Individual direct regressors only

V1.a Only one component of the individual direct regressor: booklet IDs

V1.b Only one component of the individual direct regressor: gender

V1.c Only one component of the individual direct regressor: ISEI

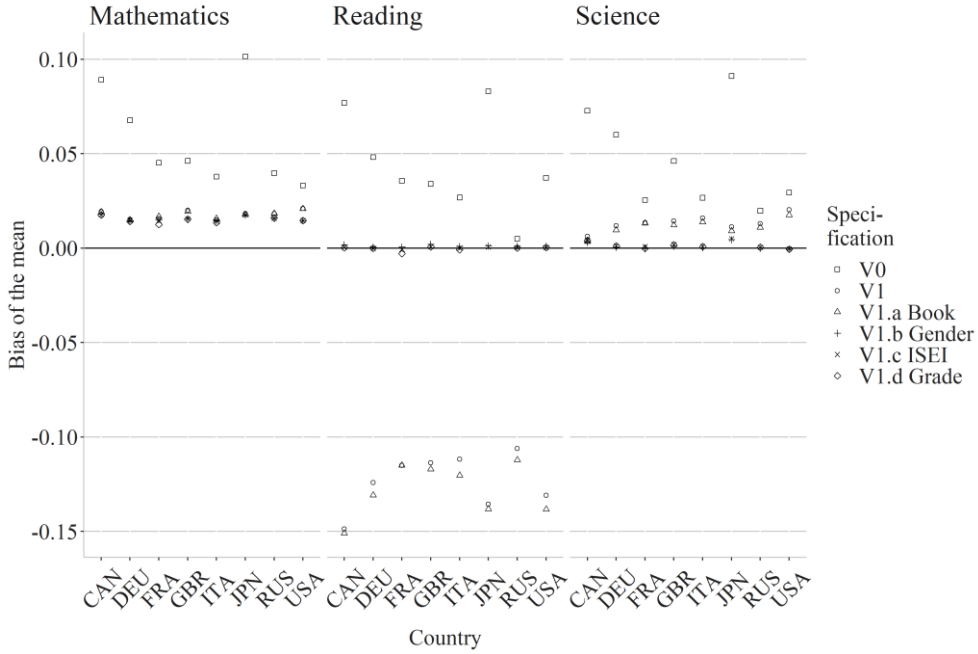
V1.d Only one component of the individual direct regressor: grade

Figure B.1 shows the bias of the mean (difference between the estimated and 'true' achievement) in standard deviations for the six different conditioning model specifications mentioned above. Thereby, the horizontal line highlights the optimal case where no bias is present. In all three domains, the bias when no conditioning is used (square) is roughly the same size and distinct from the magnitude of bias as soon as conditioning is used. For mathematics, the symbols for all conditioning specification are on top of each other meaning that there is no difference in bias between them.

This story changes dramatically when looking at reading. While most of the symbols are also close to each other and the horizontal line, the two model specifications including booklet IDs (circle: all direct regressors, triangle:

booklet IDs only) stand out. The bias increased to more than 0.1 of standard deviation and therefore, the results are more biased than no conditioning. The component of the direct regressors which introduces bias to the results is booklet IDs. Similar results, but only in a fraction of the magnitude, can be found for science.

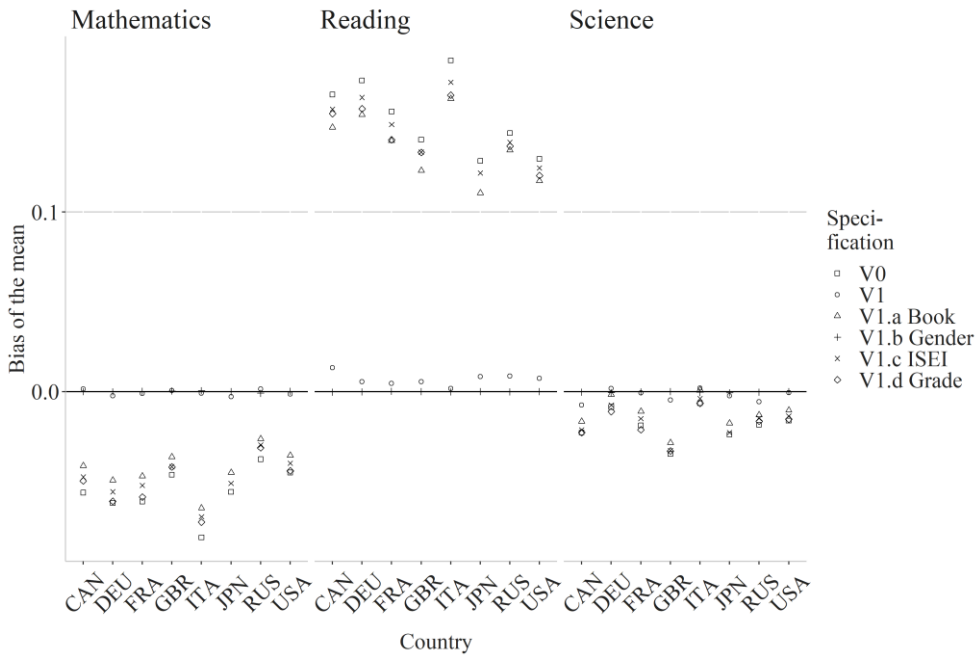
Figure B.1 Average bias of country averages in standard deviation for no conditioning, direct regressors only and the single components of the direct regressors



Note: V0: No conditioning (square), V1: Direct regressors only (circle), V1.a: Booklet ID only (triangle), V1.b: Gender only (cross), V1.c: ISEI only (x), V1.d: Grade only (diamond).

The key aspect of conditioning is counteracting the attenuation of group differences. In line with theory, I found in the main body that bias in group differences was reduced significantly as soon as the group indicator (in this case gender) was used in the model specification. This is confirmed by Figure B.2 showing the impact of the single components on the gender gaps. It becomes clear that bias is basically non-existent if only gender (cross: V1.b) is used. Direct regressors only (circle: V1) also includes gender and leads to similarly low levels of bias; only in reading is there slightly more. The other single components, which are unrelated to gender, have little to no impact on the gender gaps and stay in the range of no conditioning at all (square).

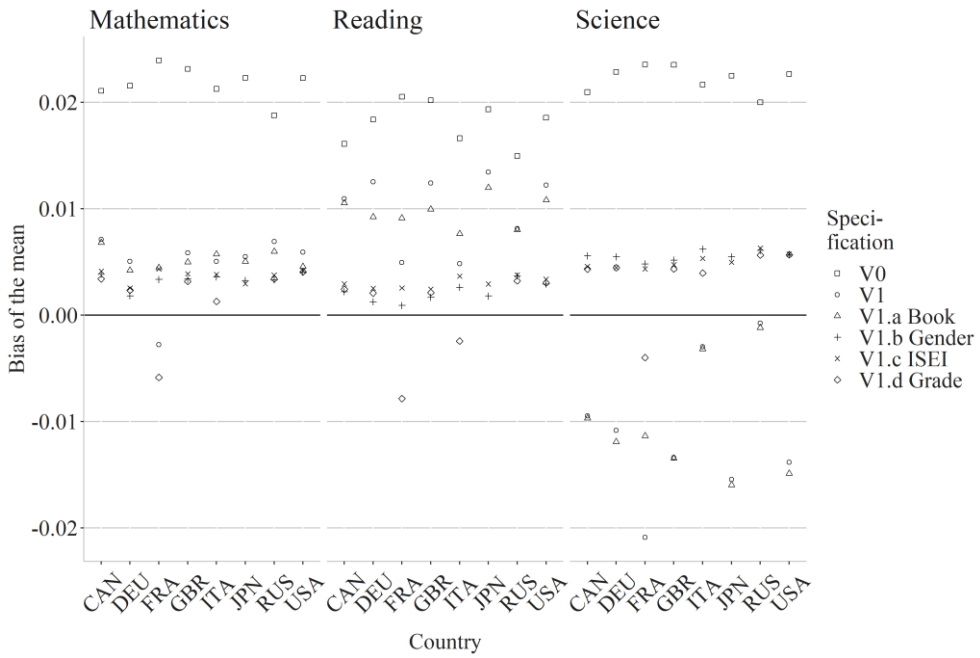
Figure B.2 Average bias of gender gaps in standard deviation for no conditioning, direct regressors only and the single components of the direct regressors



Note: V0: No conditioning (square), V1: Direct regressors only (circle), V1.a: Booklet ID only (triangle), V1.b: Gender only (cross), V1.c: ISEI only (x), V1.d: Grade only (diamond).

Figure B.3 displays the average amount of bias in the percentile differences across the six model specifications. At first glance, it seems like there is large variation in bias between the different conditioning models, but the scale needs to be taken into account. Both plots of country averages and gender gaps contain values substantially above |0.1| standard deviation, whereas the values for the percentile differences are less than |0.025| even in the extremes. In general, bias is reduced as soon as conditioning is applied. For the minor domains, reading and science, the model specifications including booklet IDs (circle: V1, triangle: V1.a) show the highest amount of bias apart from no conditioning (square), with very few exceptions.

Figure B.3 Average bias of the 90th–10th percentile differences in standard deviation for no conditioning, direct regressors only and the single components of the direct regressors



Note: V0: No conditioning (square), V1: Direct regressors only (circle), V1.a: Booklet ID only (triangle), V1.b: Gender only (cross), V1.c: ISEI only (x), V1.d: Grade only (diamond).

B.2 How would the results look if booklet IDs were excluded from the latent regression completely?

The first part of Appendix B ('B.1 Does a single component of the direct regressors introduce bias into the results?') showed that booklet IDs can introduce bias to the results in the minor domains, especially the country averages in reading. This leads to the question of whether it would be better to remove the corresponding conditioning variables altogether. In PISA 2012, the booklet IDs were deviation contrast coded and the coefficients of booklets which only contained two domains were set to zero in the latent regression for the third domain. In order to investigate whether booklet IDs should be included in the conditioning model specification and in what form, I conducted two additional models with all direct and indirect regressors (a) including booklet IDs but without restraints in the latent regression coefficients and (b) excluding booklet IDs completely.

V0. No conditioning variables (i.e. no conditioning model at all)

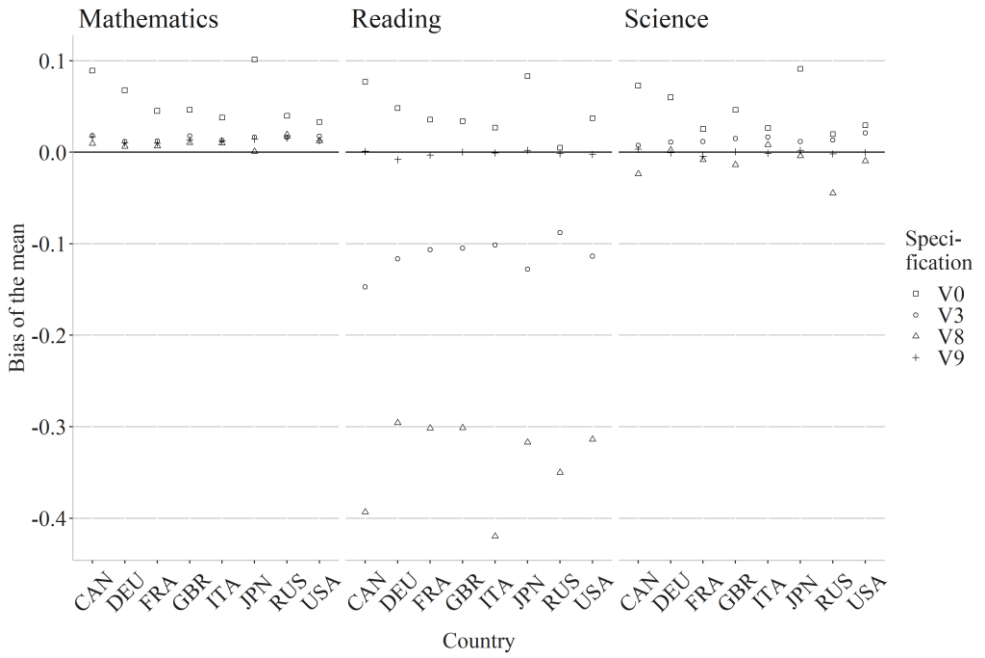
V3. Individual direct regressors and indirect regressors (all regressors)

V8. Individual direct regressors and indirect regressors (all regressors) with booklet IDs but without constraints for booklet IDs in latent regression

V9. Individual direct regressors and indirect regressors (all regressors) without booklet IDs

Figure B.4 displays the mean of the country average PISA scores when using model specifications (with different treatments of booklet IDs). This highlights how one variable can introduce severe bias into the achievement scores through conditioning. This affects the major domain only marginally, but the impact on the minor domains is substantial. When using booklet IDs without any restriction (triangle: V8), bias increases in comparison to most or all other specifications. In the case of reading, bias more than tripled in comparison to no conditioning. When booklet ID is excluded altogether (cross: V9), bias reaches its lowest level close to zero for the minor domains. The version closest to PISA 2012 is V3 which uses booklet IDs but enforces restrictions in the latent regression and can be classified in the middle ground between the two model specifications. While it does not perform as well as the model specification excluding booklet IDs, it most outperforms the model specification with booklet IDs and without latent regression coefficient restrictions, in reading even substantially.

Figure B.4 Average bias of country averages in standard deviation for variations of in-/excluding booklet IDs

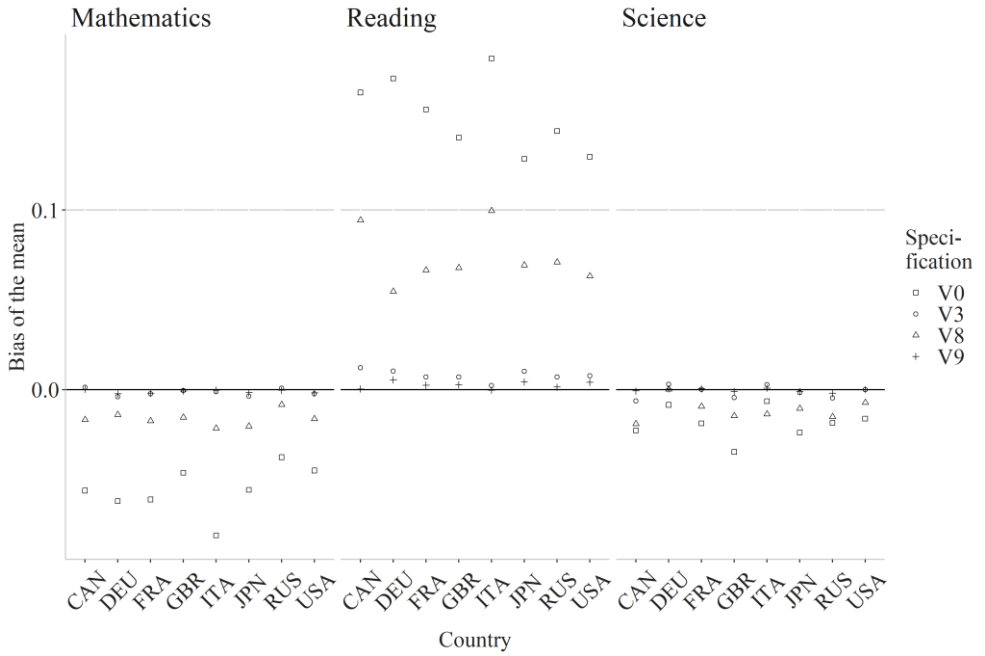


Note: V0: No conditioning at all (square), V3: Direct and indirect regressors as in PISA with booklet IDs and with constraints for booklet ID regression coefficients (circle), V8: Direct and indirect regressors with booklet IDs and without constraints for booklet ID regression coefficients (triangle), V9: Direct and indirect regressors without booklet ID (cross).

So far, the noticeable impact of booklet IDs was found for the minor domains reading and science, Figure B.5, which displays the average bias in gender gaps for the different treatments of booklet IDs, and shows that single variables can introduce bias in all domains, even the major domain. In general, bias in gender gaps decreases when conditioning. But there remain visible differences between the different model specifications. Both V3 (circle: direct and indirect regressors with booklet IDs and latent regression coefficient restrictions) and V9 (cross: direct and indirect regressors excluding booklet IDs) sit close to each other and the horizontal line denoting no bias. Thereby, V3 (circle) tends to be slightly more biased than V9 (cross) but the differences are comparably small. V8 (triangle: direct and indirect regressors with booklet IDs and without latent regression coefficient restrictions) on the other hand shows visibly more bias than V3 (circle) and V9 (cross). This highlights two points: (a) variables can introduce bias in all domains and (b) group comparisons can be biased even though the group

indicator is included in the conditioning model. In the extreme case of reading, the gender gap was biased by 0.1 standard deviation (or roughly 10 PISA points) even though gender was in the conditioning model.

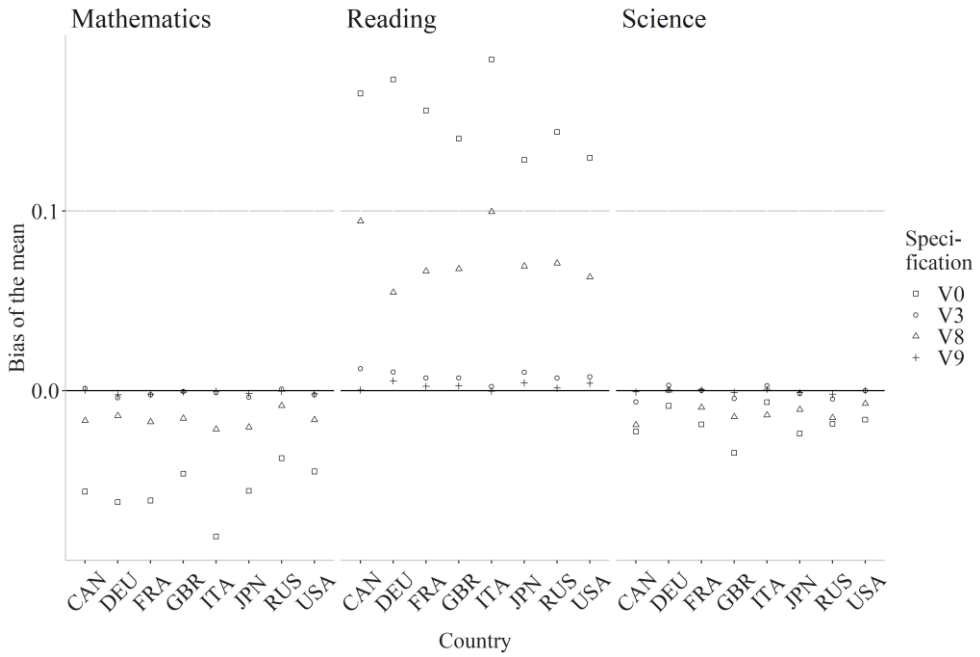
Figure B.5 Average bias of gender gaps in standard deviation for variations of in-/excluding booklet IDs



Note: V0: No conditioning at all, V3: Direct and indirect regressors as in PISA with booklet IDs and with constraints for booklet ID regression coefficients, V8: Direct and indirect regressors with booklet IDs and without constraints for booklet ID regression coefficients, V9: Direct and indirect regressors without booklet ID.

Figure B.6 displays the average bias in the percentile differences for the four conditioning model specifications. Thereby, it confirms the implications of the previous graph. While the magnitude of the bias is smaller for the percentile differences, booklet IDs also have an impact on all three domains. Bias is always largest – even larger than no conditioning (square: V0) – for the conditioning model specification with booklet IDs and without restriction (triangle: V8). While the difference between excluding booklet IDs completely (cross: V9) and using booklet IDs with latent regression coefficient restrictions (circle: V3) is minor for mathematics, V9 outperforms V3 visibly for reading and science. But again, the impact on the percentile differences is small in comparison to the other two measures.

Figure B.6 Average bias of the 90th–10th percentile differences in standard deviation for variations of in-/excluding booklet IDs



Note: V0: No conditioning at all, V3: Direct and indirect regressors as in PISA with booklet IDs and with constraints for booklet ID regression coefficients, V8: Direct and indirect regressors with booklet IDs and without constraints for booklet ID regression coefficients, V9: Direct and indirect regressors without booklet ID.

C Appendix Chapter 4

C.1 Highest parental education

Table C.1 shows all information about parental education that can be found in the official downloadable PISA 2012 data (datasets, questionnaires and codebooks downloadable from https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm). Thereby, the national versions of the questionnaires usually contain (slightly) altered versions (e.g. national qualifications are inserted instead of ISCED levels) about parental education and are already recoded to fit into the ISCED categories in the official dataset. While there are questions about parental schooling and professional education in the student background questionnaire of the official dataset, the parental questionnaire only includes questions about their professional education. As a result, more information is available from the students' responses and the corresponding composite measure is more detailed than the equivalent based on the parental responses.

Table C.1 Parental education: Questions and their response options available in the publicly available PISA database

Question-naire	Topic	Question	Response categories
Student	Schooling	What is the <highest level of schooling> completed by your mother/father?*	1 – <ISCED level 3A>, 2 – <ISCED level 3B, 3C>, 3 – <ISCED level 2 >, 4 – <ISCED level 1>, 5 – She/he did not complete <ISCED level 1>
Student	Professional education	Does your mother/father have any of the following qualifications?*	
		<ISCED level 6>	1 – Yes, 2 – No
		<ISCED level 5A>	1 – Yes, 2 – No
		<ISCED level 5B>	1 – Yes, 2 – No
		<ISCED level 4>	1 – Yes, 2 – No
Parent	Professional education	Does the child's mother/father have any of the following qualifications?	
		<ISCED 5A, 6>	1 – Yes, 2 – No
		<ISCED 5B>	1 – Yes, 2 – No
		<ISCED 4>	1 – Yes, 2 – No
		<ISCED 3A>	1 – Yes, 2 – No
Student	Composite measure	Highest educational level of parents	0 – None, 1 – ISCED 1, 2 – ISCED 2, 3 – ISCED 3B, 3C, 4 – ISCED 3A, 4, 5 – ISCED 5B, 6 – ISCED 5A, 6
Parent	Composite measure	Highest educational level of parents	0 – None, 1 – ISCED 3A, 2 – ISCED 4, 3 – ISCED 5B, 4 – ISCED 5A, 6

Note: All questions included the instruction to only tick one box. Questions with an asterisk also included the note to ask the <test administrator> for help if unsure how to answer it. The angle <brackets> denote words that needed to be inserted according to the country's equivalent (e.g. primary school for <ISCED 1> in most countries). ISCED levels correspond to: 1 – Primary education, 2 – Lower secondary education, 3 – Upper secondary education, 4 – Post-secondary non-tertiary education, 5 – First stage of tertiary education,

6 – Second stage of tertiary education. The letter after the number depends on to which further level of education it grants you access. The questions and response categories are taken from the PISA 2012 student background questionnaire codebook (OECD, n.d.-a).

The questions and response categories in the more detailed German version of the data, which is available upon request/application, are shown in Table C.2. Thereby, I translated the questions into English, but kept the response categories in German, as the German educational system is quite complex and mostly there are no equivalent well-established terms in English. I added a column 'ISCED 97 assignment' which contains the appropriate ISCED assignment for the German qualification – explanations about the German system, the terms and assignments can be found in Schneider (2008). The last column contains the coding instruction of the German qualifications onto the used categories in the official PISA 2012 dataset according to the German scaling manual/codebook (Mang et al., 2018, pp. 173–176). As can be seen in the table, the German student background and parental questionnaire contain the same questions about parental education.

Table C.2 Parental education: Question and response categories in German data as well as their mapping to ISCED 1997 scale and the PISA 2012 categories.

Student background and parental questionnaire: Schooling				
	Response categories		**ISCED 97 assignment** (Schneider, 2008)	**PISA assignment** (Mang et al., 2018)
What is the highest school-leaving qualification completed by your mother/father? *	1	Hochschulreife/ Fachhochschulreife/ Abitur	ISCED 3A	ISCED 3A, 4
	2	Berufsgrundbildungsjahr/ Berufsschule/ Berufsfachschule	Depends: ISCED 3A, 3B, 4	ISCED 3A, 4
	3	Mittlere Reife/ Realschulabschluss/ Abschluss der polytechnischen Oberschule nach der 10. Klasse (Mittlerer Abschluss)	ISCED 2	ISCED 2
	4	Hauptschulabschluss/ Volksschulabschluss	ISCED 2	ISCED 2
	5	Abschluss der Polytechnischen Oberschule nach der 8. Klasse	ISCED 2	ISCED 2
	6	Abschluss einer Sonderschule/ Förderschule	Depends – special needs school	ISCED 1
	7	Sonstiger Schulabschluss (z. B. im Ausland)	Other (e.g. abroad)	-
	8	Sie/er ist ohne Abschluss von der Schule abgegangen.	Left school without graduation	ISCED 0
	9	Sie/er hat keine Schule besucht	Did not go to school	ISCED 0
Student background and parental questionnaire: Professional education				
	Response categories		**ISCED 97 assignment**	**PISA assignment**
	y/n	Promotion (Doktorprüfung)	ISCED 6	ISCED 6, 5A

Does your mother/father	y/n	Hochschulabschluss (Magister/ Diplom/ Staatsexamen/ Bachelor/ Master)	ISCED 5A	ISCED 6, 5A
have one or	y/n	Fachhochschulabschluss/Diplom (FH)	ISCED 5A	ISCED 6, 5A
multiple of the	y/n	Abschluss an einer Fachschule/ Meister- oder Technikerschule/ einer Schule des Gesundheitswesens/ Abschluss an einer Berufsakademie/ Fachakademie (oder ein vergleichbarer Abschluss im Ausland)	ISCED 5B	ISCED 5B
following qualifications? *	y/n	Abschluss an einer Fachoberschule/ Berufsschule/ Berufsfachschule/ Berufsoberschule/ Technische Oberschule (oder ein vergleichbarer Abschluss im Ausland)	ISCED 4	ISCED 3A, 4
	y/n	Abgeschlossene Lehre (Abschluss an einer Handelsschule oder ein vergleichbarer Abschluss im Ausland)	Depends – ISCED 3B, C	ISCED 3B, 3C
	y/n	Sonstiger beruflicher Abschluss (z.B. im Ausland)	Other (e.g. abroad)	-

Note: Response categories are kept in German, as some of them have no equivalent in English, see Schneider (2008) for an overview of the German educational system and explanations of their equivalent/educational levels and intents. Question with an asterisk also included the note to ask the <test administrator> for help if unsure how to answer it. For the first question, the highest qualification should be ticked. The second question has y/n (yes/no) options and every obtained qualification should be indicated. ISCED levels correspond to: 1 – Primary education, 2 – Lower secondary education, 3 – Upper secondary education, 4 – Post-secondary non-tertiary education, 5 – First stage of tertiary education, 6 – Second stage of tertiary education. The letter after the number depends on to which further level of education it grants you access. The questions (in German – I translated them into English) are taken from the German scale manual/codebook (Mang et al., 2018).

C.2 Technical details of the plausible value computation in Chapter 4

The plausible value computation is implemented in Chapter 4 as follows:

- The scored cognitive item response data from PISA 2012, downloadable from https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm, and the national student background questionnaire data for Germany is used (Prenzel et al., 2015; available after application) are used in the plausible value computation.
- Only one factor variable is used in two of the three conditioning models: highest parental education. The third does not use any conditioning at all. Highest parental education is computed based on the national student background questionnaire data according to the planned assignment (Mang et al., 2018; also see Subchapter 4.4.3). Two versions of highest parental education are computed: one based on the students' responses and one based on the parental responses.
- The item response theory model uses the official reported item difficulties as fixed item difficulties in the computation (OECD, 2014b, pp. 406–413) with one exception. I believe that the difficulty of PM155Q03D is a typo. As a result, I used the average value from all previous cycles ($\tau_1 = 0.184, \tau_2 = -0.184$).
- The conditioning model is conducted using a 'divide-and-conquer' approach (Patz & Junker, 1999; van Rijn, 2018). This means that the IRT model is computed first and the latent regression subsequently.
- The IRT model is estimated using `tam.mml()` and latent regression using `tam.latreg()` from the R package 'TAM' (Robitzsch et al., 2018). Further computational details are: quasi-Monte Carlo integration, 2000 nodes, convergence criterium for deviance=0.001, convergence criteria for coefficients=0.0001.
- Three version are computed:
    1. No conditioning at all (only the IRT model)

2. Conditioning: only using highest parental education based on the students' responses
3. Conditioning: only using highest parental education based on the parental responses

- Five plausible values are drawn from the resulting multivariate normal distributions, each with the help of `tam.pv()` (Robitzsch et al., 2018). Further computational details are: Monte Carlo estimation, 2000 nodes.

Bibliography

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education*, *1*(5). https://doi.org/10.1186/2196-0739-1-5

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Addey, C. (2015). Participating in international literacy assessments in Lao PDR and Mongolia: A global ritual of belonging. In B. Hamilton, C. Maddox, & C. Addey (Eds.), *Literacy as numbers: Researching the politics and practices of international literacy assessment* (pp. 147–164). Cambridge University Press.

Addey, C. (2017). Golden relics & historical standards: How the OECD is expanding global education governance through PISA for Development. *Critical Studies in Education*, *58*(3), 311–325. https://doi.org/10.1080/17508487.2017.1352006

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger, A. (2017). The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*, *47*(3), 434–452.

Anders, J., Has, S., Jerrim, J., Shure, N., & Zieger, L. (2021). Is Canada really an education superpower? The impact of non-participation on results from PISA 2015. *Educational Assessment, Evaluation and*

Accountability, *33*, 229–249. https://doi.org/10.1007/s11092-020-09329-5

Auld, E., & Morris, P. (2019). The OECD and IELS: Redefining early childhood education for the 21st century. *Policy Futures in Education*, *17*(1), 11–26. https://doi.org/10.1177/1478210318823949

Avvisati, F., Le Donné, N., & Paccagnella, M. (2019). A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences*, *1*(8). https://doi.org/10.1186/s42409-019-0010-z

Behr, D., & Zabal, A. (2019). A meeting report: OECD-GESIS seminar on translating and adapting instruments in large-scale assessments (2018). *Measurement Instruments for the Social Sciences*, *1*(10). https://doi.org/10.1186/s42409-019-0011-y

Billiet, J., & Matsuo, H. (2012). Non-response and measurement error. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 149–178). Springer New York. https://doi.org/10.1007/978-1-4614-3876-2_10

Bloem, S. (2015). The OECD directorate for education as an independent knowledge producer through PISA. In H.-G. Kotthoff & E. Klerides (Eds.), *Governing educational spaces* (pp. 169–185). Brill Sense.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers, No 71*. https://doi.org/10.1787/5k9fdfqffr28-en

Breen, R., & Jonsson, J. O. (2005). Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annual Review of Sociology*, *31*, 223–243.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292. https://doi.org/10.1002/sim.2673

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

Caro, D. H., & Cortés, D. (2012). Measuring family socioeconomic status: An illustration using data from PIRLS 2006. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, *5*, 9–33.

Carpenter, H., Papps, I., Bragg, J., Dyson, A., Harris, D., Kerr, K., Todd, L., & Laing, K. (2013). *Evaluation of pupil premium: Research brief*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/210063/DFE-RB282.pdf

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2014). Multivariate analysis of mixed data: The PCAmixdata R package. *ArXiv Preprint ArXiv:1411.4911*, *132*.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*(2), 294–304. https://doi.org/10.1037/0893-3200.19.2.294

Davoli, M., & Entorf, H. (2018). The PISA shock, socioeconomic inequality, and school reforms in Germany. *IZA Policy Paper*.

Education Datalab. (2017). *Why does Vietnam do so well in PISA? An example of why naive interpretation of international rankings is such a bad idea*. https://ffteducationdatalab.org.uk/2017/07/why-does-vietnam-do-so-well-in-pisa-an-example-of-why-naive-interpretation-of-international-rankings-is-such-a-bad-idea/

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.

Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy*, *34*(4), 470–499.

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

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

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

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

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

Gromada, A., Rees, G., Chzhen, Y., Cuesta, J., & Bruckauf, Z. (2018). Measuring inequality in children's education in rich countries. *Innocenti Working Papers*, *18*. https://doi.org/10.18356/5f90f95e-en

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879.

Guryan, J., Hurst, E., & Kearney, M. (2008). Parental education and parental time with children. *Journal of Economic Perspectives*, *22*(3), 23–46. https://doi.org/10.1257/jep.22.3.23

Hansen, K. Y., & Gustafsson, J.-E. (2016). Determinants of country differences in effects of parental education on children's academic achievement. *Large-Scale Assessments in Education*, *4*(11).

Heine, J.-H., Nagy, G., Meinck, S., Zühlke, O., & Mang, J. (2017). Empirische Grundlage, Stichprobenausfall und Adjustierung im PISA-Längsschnitt 2012–2013. *Zeitschrift Für Erziehungswissenschaft*, *20*(2), 287–306. https://doi.org/10.1007/s11618-017-0756-0

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353.

Hopkins, D., Pennock, D., Ritzen, J., Ahtaridou, E., & Zimmer, K. (2008). *External evaluation of the policy impact of PISA (Vol. EDU/PISA/GB(2008)35/REV1)*. OECD. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB(2008)35/REV1&docLanguage=En

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

Howie, S., & Plomp, T. (2005). International comparative studies of education and large-scale change. In N. Bascia, A. Cumming, A. Datnow, K. Leithwood, & D. Livingstone (Eds.), *International handbook of educational policy* (Vol. 1, pp. 75–99). Springer.

Husén, T., & Postlethwaite, T. N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education, 3*(2), 129–141.

Jerrim, J. (2013). *The Reading Gap: The socio-economic gap in children's reading skills: A cross-national comparison using PISA 2009*. The Sutton Trust.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review, 61*, 51–58. https://doi.org/10.1016/j.econedurev.2017.09.007

Jerrim, J., & Macmillan, L. (2015). Income inequality, intergenerational mobility, and the great Gatsby curve: Is education the key? *Social Forces*, *94*(2), 505–533.

Jerrim, J., & Micklewright, J. (2014). Socio-economic gradients in children's cognitive skills: Are cross-country comparisons robust to who reports family background? *European Sociological Review*, *30*(6), 766–781. https://doi.org/10.1093/esr/jcu072

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

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research*, *58*, 139–148. https://doi.org/10.1080/00131881.2016.1165559

Kaiser, L. C. (2018). Das Bildungs-und Teilhabepaket: Eine Miss-/Erfolgsstory? In L. C. Kaiser (Ed.), *Soziale Sicherung im Umbruch* (pp. 145–161). Springer VS. https://doi.org/10.1007/978-3-658-06502-7_7

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

Karsten, S. (2006). Policies for disadvantaged children under scrutiny: The Dutch policy compared with policies in France, England, Flanders and the USA. *Comparative Education*, *42*(02), 261–282.

Khorramdel, L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2020). Plausible values: Principles of item response theory and multiple imputations. In D. B. Maehler & B. Rammstedt (Eds.), *Large-Scale Cognitive Assessment: Analyzing PIAAC Data* (pp. 27–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_3

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

LeRoy, B. W., Samuel, P., Deluca, M., & Evans, P. (2019). Students with special educational needs within PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 386–396.

Lewis, S. (2020). *PISA, policy and the OECD: Respatialising global educational governance through PISA for Schools*. Springer. 10.1007/978-981-15-8285-1

Lohmann, H., Spieß, C. K., & Feldhaus, C. (2009). Der Trend zur Privatschule geht an bildungsfernen Eltern vorbei. *DIW Wochenberich, 76,* 640–646.

Ludeke, S. G., Gensowski, M., Junge, S. Y., Kirkpatrick, R. M., John, O. P., & Andersen, S. C. (2021). Does parental education influence child educational outcomes? A developmental analysis in a full-population sample and adoptee design. *Journal of Personality and Social Psychology*, *120*(4), 1074–1090.

Mang, J., Ustjanzew, N., Schiepe-Tiska, A., Prenzel, M., Sälzer, C., Müller, K., & González Rodríguez, E. (2018). *PISA 2012 Skalenhandbuch. Dokumentation der Erhebungsinstrumente*. Waxmann.

Martin, M. O., Mullis, I. V., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martins, L., & Veiga, P. (2010). Do inequalities in parents' education play an important role in PISA students' mathematics achievement test score disparities? *Economics of Education Review, 29*(6), 1016–1033. https://doi.org/10.1016/j.econedurev.2010.05.001

Matta, T. H., Rutkowski, L., Rutkowski, D., & Liaw, Y.-L. (2018). lsasim: An R package for simulating large-scale assessment data. *Large-Scale Assessments in Education*, *6*(1), 1–33. https://doi.org/10.1186/s40536-018-0068-8

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.

Molnar, M. (2014, December 16). *Pearson Wins Bid to Develop PISA 2018 Frameworks*. https://marketbrief.edweek.org/marketplace-k-12/pearson_wins_bid_to_develop_pisa_2018_frameworks/

Morris, T., White, I., & Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102.

Mullis, I. V. S., Martin, M., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-results/

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/pirls2016/international-results/

OECD. (n.d.-a). *Codebook for PISA 2012 main study student questionnaire*. https://www.oecd.org/pisa/pisaproducts/PISA12_stu_codebook.pdf

OECD. (n.d.-b). *Programme for International Student Assessment 2012 (PISA 2012) [Data set]*. Retrieved November 18, 2020, from https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm

OECD. (1999). *Classifying educational programmes: Manual for ISCED-97 implementation in OECD countries*. OECD Publishing.

OECD. (2013). *PISA 2012 results: Excellence through equity (Volume II)*. OECD Publishing. http://dx.doi.org/10.1787/9789264201132-en

OECD. (2014a). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science* (Volume I, Revised edition, February 2014). OECD Publishing. https://doi.org/10.1787/9789264208780-en

OECD. (2014b). *PISA 2012 technical report*. OECD Publishing.

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. https://www.oecd-ilibrary.org/content/publication/b25efab8-en

OECD. (2019b). *PISA 2018 results (Volume I)*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD. (2019c). *PISA 2018 results (Volume II): Where all studentes can succeed*. OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en

OECD. (2019d). *Relationship between the Survey of Adult Skills (PIAAC) and the OECD Programme for International Student Assessment (PISA)*. https://www.oecd-ilibrary.org/content/component/4643faea-en

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

Pishghadam, R., & Zabihi, R. (2011). Parental education and social and cultural capital in academic achievement. *International Journal of English Linguistics*, *1*(2), 50–57.

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland*. Waxmann.

Prenzel, M., Sälzer, C., Klieme, E., Köller, O., Mang, J., Heine, J.-H., Schiepe-Tiska, A., & Müller, K. (2015). *Programme for International*

Student Assessment 2012 (PISA 2012) (Version 5) [Data set]. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2012_v5

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

Ridolfo, H., & Maitland, A. (2011). Factors that influence the accuracy of adolescent proxy reporting of parental characteristics: A research note. *Journal of Adolescence*, *34*(1), 95–103.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

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

Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–153). Chapman and Hall/CRC Press.

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education*, *8*(3), 259–278. https://doi.org/10.2304/rcie.2013.8.3.259

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, *48*(3), 293–312. https://doi.org/10.1111/j.1745-3984.2011.00144.x

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. (2017). Design considerations for planned missing auxiliary data in a latent regression context. *Psychological Test and Assessment Modeling*, *59*(1), 55–70.

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430. https://doi.org/10.1080/00220272.2010.487546

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

Schleicher, A. (2000). Monitoring student knowledge and skills: The OECD Programme for International Student Assessment. In D. Shorrocks-Taylor & E. W. Jenkins (Eds.), *Learning from others: International comparisons in education* (pp. 63–77). Springer.

Schleicher, A. (2013). *Use data to build better schools*. TEDGlobal. http://www.ted.com/talks/andreas_schleicher_use_data_to_ build_better_schools?language=en

Schleicher, A. (2019). *PISA 2018: Insights and interpretations.* OECD Publishing.

Schneider, S. L. (2008). Applying the ISCED-97 to the German educational qualifications. *The International Standard Classification of Education*

(ISCED-97). An Evaluation of Content and Criterion Validity For, *15*, 76–102.

Schneider, S. L., & Kogan, I. (2008). *The International Standard Classification of Education 1997: Challenges in the application to national data and the implementation in cross-national surveys*. MZES.

Schnepf, S. V. (2018). *Insights into survey errors of large scale educational achievement surveys* (No. JRC111734; JRC Working Papers in Economics and Finance). Publications Office of the European Union.

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

Sjøberg, S. (2017). Pisa testing. *Europhysics News*, *48*(4), 17–20. https://doi.org/10.1051/epn/2017402

Socio-Economic Panel. (2019). *Socio-economic panel (SOEP): Data for years 1984–2018, version 35 [Data set]*. SOEP. doi:10.5684/soep.v35

Statistische Ämter des Bundes und der Länder. (2011). *Sozioökonomische Grunddaten für die zwölf neuen EU-Mitgliedsstaaten, Deutschland und die Bundesländer*. Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen. https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Publikationen/Downloads/soziooekonomische-grunddaten.html

Statistisches Bundesamt. (2012). *Statistisches Jahrbuch: Deutschland und Internationales*. Statistisches Bundesamt. https://www.statistischebibliothek.de/mir/receive/DEAusgabe_mods_00000380

Steiner-Khamsi, G. (2003). The politics of league tables. *Journal of Social Science Education*. https://doi.org/10.4119/jsse-301

Steiner-Khamsi, G., & Stolpe, I. (2006). *Educational import: Local encounters with global forces in Mongolia*. Palgrave Macmillan.

Steiner-Khamsi, G., & Waldow, F. (Eds.). (2012). *World yearbook of education 2012: Policy borrowing and lending in education*. Routledge.

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

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

Volante, L., Schnepf, S. V., Jerrim, J., & Klinger, D. A. (2019). *Socioeconomic inequality and student outcomes: Cross-national trends, policies, and practices* (Vol. 4). Springer.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski, *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). Chapman Hall/CRC.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

von Davier, M., Rutkowski, D., & Rutkowski, L. (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, Fla.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education*, *58*(3), 152–166.

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

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

Yeung, W. J., Linver, M. R., & Brooks–Gunn, J. (2002). How money matters for young children's development: Parental investment and family processes. *Child Development*, *73*(6), 1861–1879.

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, *21*, 1–22.

Zieger, L. (2021). *Code for "Conditioning: How background variables can influence PISA scores."* osf.io/8fzns